University of Alberta

Biochemical Applications of Molecular Modeling and Docking

by

Maxwell David Cummings ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Department of Biochemistry

Edmonton, Alberta

Fall 1996

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-612-18026-3

Canada

University of Alberta

Library Release Form

Name of Author: Maxwell David Cummings

Title of Thesis: Biochemical Applications of Molecular Modeling and Docking

Degree: Doctor of Philosophy

Year this Degree Granted: 1996

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Max Cummings
10820 73 Avenue
Edmonton, Alberta, Canada
T6E 1C7

20/8/96
Date

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Biochemical Applications of Molecular Modeling and Docking submitted by Maxwell David Cummings in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Randy Read
Randy J. Read (supervisor)

David Bundle
David R. Bundle

Michael J. Ellison

Michael James
Michael N.G. James

Brian K. Shoichet (external)

Joel Weiner
Joel H. Weiner

7 August 1996
Date

CAMBRIDGE
UNIVERSITY PRESS

North American Branch
40 West 20th Street
New York, NY 10011-4211
USA

Telephone 212 924 3900
Fax 212 691 3239

PERMISSION TO REPRODUCE IN DISSERTATION

July 12, 1996

Max Cummings
MMID, 1-41 Medical Sciences Bldg.
University of Alberta
Edmonton, Alberta T6G 2H7
Canada

Reference

ISBN/Journal:	*Protein Science*, Vol. 4 (1995), pp. 2087-2099
Author:	Cummings, M.D., Hart, T.N. and Read, R.J.
Title:	"Atomic Solvation Parameters in the Analysis of Protein-Protein Docking Results"
Item/pp.:	excerpts as needed

Use

University/College: University of Alberta

Rights/Acknowledgement

Permission is granted for non-profit educational use in your dissertation, subject to full acknowledgement of the material and clear indication of the copyright notice as it appears in our publication, followed by the phrase "Reprinted with the permission of Cambridge University Press."

Restrictions

This permission does not allow reprinting any material copyrighted by or credited in our publication to another source; Cambridge disclaims all liability in connection with the use of such material without proper consent.

Authorization: ______________________
M.P. Anderson
Rights and Permissions Manager

Abstract

The development of general strategies for the performance of docking simulations is prerequisite to the exploitation of this powerful computational method. For both manual and automated docking, the development of comprehensive and reliable strategies depends upon experience with a diverse array of problems. Here we report modeling and docking studies aimed at general strategy development (Chapters 2 and 3), as well as investigations of specific biochemical questions (Chapters 4 and 5).

Automated docking was used to reconstruct diubiquitin from its two halves, as well as from two copies of the uncomplexed monomer. The correct solutions were ranked amongst the most favorable in all of the systems studied. In the experiments involving the ubiquitin monomer, various structural modifications were made to compensate for the lack of flexibility and for the lack of a covalent bond in the modeled interaction; a variety of analyses was performed on the low energy dockings obtained in these experiments. Characterization of the interface surfaces, as well as mechanistic information, enabled us to distinguish more accurately between correct and incorrect dockings.

Our initial studies with ubiquitin led us to investigate more thoroughly the use of atomic solvation parameters (ASPs) to approximate bulk desolvation in protein-protein docking. We re-derived nine different ASP sets from literature data, and chose three for further testing. For most of the docking results we analyzed, the use of an octanol-water-based ASP set marginally improved the energetic ranking of the low energy dockings, whereas the other ASP sets we tested disturbed the ranking of the low energy dockings in many of the same systems. A similar conclusion was reached when we examined the correlation between the experimental and calculated interaction energies for a series of proteinase-inhibitor complexes.

We modeled three unique sites for binding of the carbohydrate moiety of globotriaosylceramide (Gb3) to the wild-type binding pentamer of pig edema toxin

(SLT-IIe) and the double mutant GT3, based upon the three sites observed for the related verotoxin-1 binding pentamer. Examination of the three sites in light of various mutation and binding data strongly suggested one binding site, in preference over the other two. We applied several modeling techniques, and developed a model for binding of the carbohydrate moiety of globotetraosylceramide (Gb4) to this site of the SLT-IIe binding pentamer. This model is consistent with a wide variety of mutation and binding data, and clearly shows the importance of the terminal GalNAc residue of Gb4, as well as the two mutated residues of GT3, to the intermolecular interaction.

Several new flexible docking and superposition tools, as well as a more conventional rigid-body (fragment) docking method (docking methods developed in this laboratory), were used to examine NAD binding to the catalytic subunits of diphtheria (DT) and pertussis (PT) toxins, and to propose a model of the NAD-PT complex. Low energy dockings of the rigid NAD fragments adenine and nicotinamide clustered in three distinct sites on the two proteins; two of the sites were common to both fragments, and were related to the NAD-DT structure in an obvious way. However, the adenine subsite of PT was shifted relative to that of DT. We chose adenine/nicotinamide pairs of PT dockings from these clusters, and flexibly superimposed NAD onto these pairs. The lowest energy NAD-PT model accounts for the sequence and structural similarities between PT and DT, and is consistent with many results that suggest the catalytic importance of certain residues. A possible functional role for the structural difference between the two complexes is discussed.

Acknowledgements

The work described in this thesis involved protein structures and computer software from many different sources, and in almost all of these cases I had no personal involvement in producing the original data or program. My research has relied on the efforts of many other people, and I am indebted to all of them.

Randy Read provided a challenging and stimulating atmosphere for me to learn about protein modeling and docking. He let me go my own way when he probably knew better, and he kept his calm when I didn't keep mine. Randy's examples of intelligence, diplomacy, and critical analysis have been truly educational, and I will have done well if I take home only a small part of those lessons. It was also particularly helpful of him to point out that hydrogen should not be bonded to two carbons.

Trevor Hart was also a major contributor to the challenging and stimulating atmosphere in which I spent the last four years. Trevor wrote all of the docking software and most of the analysis tools that I used in my docking studies, and also provided me with a wealth of theoretical and practical aid, especially in the early stages of my work. In return, and I think it was a pretty even deal, I chased way too many wild pitches. Overall, I think we both learnt a thing or two.

Randy and Trevor, I thank you both very much for your friendship and guidance.

The other members of the lab were also a joy to work with and be around. I thank them all for the significant contribution each of them made toward the great time I have had over the past four years.

Dave Bundle, Mike Ellison, Mike James, Brian Shoichet, and Joel Weiner comprised my examining committee. Their thoughtful comments, criticisms, and suggestions were and are appreciated. Mike Ellison was also a stimulating collaborator on the ubiquitin project.

Hong Ling, David Eisenberg, and David McKay kindly provided protein structures that were not yet publicly available.

Many people, besides Randy and Trevor, helped me with the use of software: Mike Bass wrote the NETWORK program, helped me to get it running, and even provided a revised version based on a few comments I made; Stephen Evans wrote the SETOR/GRAPH program and provided me with a few sample files and hints; the people at the Biosym Response Center, particularly Marvin Waldman and Fora Chan, were very helpful in the early stages of my work; and finally, many of the people in Randy's and Mike James' labs helped me with running software.

Over the course of my undergraduate and graduate studies the Natural Sciences and Engineering and Research Council of Canada have geneously provided me with financial support, for which I am grateful. More recently SynPhar Labs and Randy Read have provided financial support, and this is also greatly appreciated.

During the past four years I have been on leave from my position at SynPhar Labs. I thank my former supervisor, Kazuo Adachi, for being so supportive of my plan to return to school, encouraging SynPhar to support me, and for being so interested during the early stages of my program. I also thank Ron Micetich, President of SynPhar, for his generous financial support over the past four years.

On a more personal note, I thank my Mom, Estelle Cummings, for her generous support, financial and otherwise, during my undergraduate studies and the earlier years of my graduate studies. Both of my parents, Estelle and David, instilled in me a healthy respect for the value of education, and I thank them both for this.

Finally, I thank Gudrun Trescher for putting up with me during all the highs and lows of the past few years, and for giving me love and understanding when I needed it.

Contents

1 Introduction **1**

1.1 Protein-ligand interactions 1

1.2 Simulating protein-ligand interactions 7

1.3 Some applications of automated docking 14

1.4 Summary of the Introduction 22

1.5 References 23

2 Monte Carlo docking with ubiquitin **27**

2.1 Introduction 27

2.2 Materials and methods 31

2.2.1 Structures 31

2.2.2 Hardware 31

2.2.3 Software 32

2.2.4 Structure preparation 32

2.2.5 Reference structures 33

2.2.6 Docking 34

2.2.7 Analysis - reconstruction of diubiquitin from its two halves . . 35

2.2.8 Analysis - construction of diubiquitin from two ubiquitin monomers 35

2.3 Results and Discussion 36

2.3.1 Relevant biochemical information 36

2.3.2 Relevant structural information 37

2.3.3 Docking - summary of experimental constraints 39

2.3.4 Docking - criteria of success 41

2.3.5 Docking - reconstruction of diubiquitin from its two halves . . 43

2.3.6 Docking - construction of diubiquitin with two copies of (mono)ubiquitin 48

2.3.7 Docking - tetraubiquitin 57

2.3.8 Docking - summary 60

2.4 Conclusion 64

2.5 References 65

3 ASPs in protein-protein docking **69**

3.1 Introduction 69

3.2 Materials and methods 74

3.2.1 Previously published data 74
3.2.2 Calculation of atomic solvation parameters 74
3.2.3 Accessible surface area calculations 76
3.2.4 Docking simulations and energy calculations 76
3.3 Results and Discussion 80
3.3.1 Development of new ASPs 80
3.3.2 Evaluation of ASPs: criteria in docking 86
3.3.3 Evaluation of ASPs: docking with SGPB-OMTKY3 88
3.3.4 Evaluation of ASPs: docking with ubiquitin 91
3.3.5 Evaluation of ASPs: a series of protease-inhibitor complexes . 95
3.3.6 Evaluation of ASPs: summary 97
3.4 Conclusion 100
3.5 References 101

4 Modeling carbohydrate-binding specificity **105**
4.1 Introduction 105
4.2 Materials and methods 107
4.2.1 Protein structure preparation 107
4.2.2 Calculation of carbohydrate conformations 109
4.2.3 Carbohydrate-protein complexes 110
4.3 Results and discussion 111
4.3.1 Calculated Gb3 carbohydrate conformations 111
4.3.2 Comparison of calculated and bound Gb3 conformations . . . 113
4.3.3 Gb3 binding sites on SLT-I and GT3 114
4.3.4 Gb4 conformations 119
4.3.5 Gb4 binding at site III of SLT-IIe 120
4.3.6 Plausibility of the model 128
4.4 References 134

5 Fragment-based modeling of NAD binding **138**
5.1 Introduction 138
5.2 Materials and methods 141
5.2.1 Preparation of docking targets and ligands 141
5.2.2 Superpositioning and rigid-body energy minimization 142
5.2.3 Docking simulations 144
5.3 Results and discussion 147
5.3.1 Preliminary analysis of relevant available complex structures . 147
5.3.2 NAD fragment docking to DT and PT 150
5.3.3 Modeling the NAD-PT complex 155
5.3.4 A note on ligand design 162
5.3.5 Significance 163
5.4 References 164

6 General Discussion and Conclusions **168**
6.1 References 178

A Methods: the BOXSEARCH docking program **179**
A.1 BOXSEARCH theory 180
A.2 BOXSEARCH practice 182
A.2.1 Probe and target preparation 183
A.2.2 The floating grid 184
A.2.3 Residue and charge group libraries 185
A.2.4 Preliminary docking simulations 186
A.2.5 Analysis of docking output 186
A.3 References 189

B Flexibility in automated docking **191**
B.1 References 196

List of Tables

2.1 The Monte Carlo minimization schedule. 33
2.2 The annealing schedule for Monte Carlo docking 34
2.3 Reference structures for docking experiments 42
2.4 Docking statistics for probe 1. 46
2.5 Top 10 clusters for large experiment with probe 1 47
2.6 Top 10 clusters for experiment with modified ubiquitin 52

3.1 Three sets of surface area data for regression analysis 79
3.2 Three sets of transfer free energies for regression analysis 82
3.3 Statistics for various transfer free energy - surface area regressions . . 84
3.4 Atomic solvation parameters derived from various regressions 85
3.5 Low energy dockings and reference configurations for the docking systems 87

4.1 Low energy Gb3 conformers 112
4.2 Glycoside conformations of bound Gb3 115
4.3 Low energy Gb4 conformers 119
4.4 Minimization constraints 121
4.5 Residue-residue interactions at site III 122
4.6 Residue-residue interactions for the final model of Gb4 at site III . . . 126
4.7 Possible hydrogen bonds of the final Gb4-site III model 127

5.1 Flexible NAD dihedrals 142
5.2 Adenine and nicotinamide dockings to DT 152
5.3 Adenine and nicotinamide dockings to PT 154
5.4 Refined models of the NAD-PT complex 157
5.5 Intermolecular contacts for dockings 1 and 2 158
5.6 Intermolecular hydrogen bonds for docking 1 160

List of Figures

1.1 Protein-ligand interactions. 2
1.2 Generalized docking algorithm. 9

2.1 Flexibility in ubiquitin and diubiquitin 38
2.2 Reconstruction of diubiquitin from its two halves 44
2.3 Superposition of the lowest energy docking for probe 1 45
2.4 Superposition of mono- and di-ubiquitin 49
2.5 Reconstruction of diubiquitin from "mutant" ubiquitin 51
2.6 Docking with monoubiquitin 54

3.1 Flowchart showing the procedure for re-analysis of docking results . . 77
3.2 Correlation of calculated and experimental transfer free energies of amino acid analogs for 3 different ASP sets 83
3.3 Energy differences for complexed SGPB-OMTKY3 89
3.4 Energy differences for native SGPB-OMTKY3 90
3.5 Energy differences for complexed SGPB-FRAG1 91
3.6 Energy differences for complexed diubiquitin 92
3.7 Energy differences for modified diubiquitin 93
3.8 Energy differences for native (mono)ubiquitin 94
3.9 Correlation of calculated interaction energies and experimental free energies for a series of proteinase-inhibitor complexes 97

4.1 Analysis of glycosidic linkages. 111
4.2 Gb3 binding sites of SLT-I and GT3. 114
4.3 Alignment of SLT-I with GT3. 117
4.4 Initial model of Gb3 binding sites of SLT-IIe 118
4.5 Minimized complexes of Gb4 conformers at site III 124
4.6 The final model of Gb4 bound at site III of SLT-IIe. 127

5.1 NAD atom names. 143
5.2 Extent of docking search. 144
5.3 Sequence alignment for superpositioning. 145
5.4 The NAD binding site of DT. 146
5.5 The NAD binding site of ETA. 148
5.6 Fragment dockings to DT 153
5.7 Fragment dockings to PT 155

5.8 Modeling the NAD-PT complex 156
5.9 Low energy models of NAD-PT 159

A.1 Spatial extents of a BOXSEARCH simulation. 184
A.2 Cluster analysis of docking output. 187

DEFINITIONS

Abbreviations

AChE	acetylcholinesterase
ADPR	adenosine 5'-diphosphate ribosyltransferase
ADP	adenosine 5'-diphosphate
AMP	adenosine 5'-monophosphate
ApUp	adenylyl 3'-5' uridine 3' monophosphate
ASP	atomic solvation parameter
ATP	adenosine 5'-triphosphate
DT	diphtheria toxin
E1	ubiquitin activating enzyme
E2	ubiquitin conjugating enzyme
E3	ubiquitin ligase
EF-2	elongation factor 2
ETA	*Pseudomonas aeruginosa* exotoxin A
FRAG1	reactive site tripeptide of OMTKY3
Gal	galactose
Gal*N*Ac	*N*-acetylgalactosamine
Glc	glucose
Gb3	αGal(1-4)βGal(1-4)βGlcCer, where Cer is ceramide
Gb4	βGal*N*Ac(1-3)αGal(1-4)βGal(1-4)βGlcCer, where Cer is ceramide
GT3	double mutant of SLT-IIe
hGH	human Growth Hormone
hGHR	human Growth Hormone Receptor
hNAD	hydrolyzed NAD (in this case nicotinamide and AMP)
NAD	nicotinamide adenine dinucleotide
OMTKY3	third domain of turkey ovomucoid inhibitor
PDB	Brookhaven Protein Data Bank
PT	pertussis toxin
PTH	phenolpthalein
SAR	structure-activity relationship
SB	sulisobenzone
SGPB	*Streptomyces griseus* Proteinase B
SLT-I	Shiga-like toxin I (verotoxin 1, VT-1)
SLT-II	Shiga-like toxin II (verotoxin 2, VT-2)
SLT-IIc	Shiga-like toxin IIc (verotoxin 2c, VT-2c)
SLT-IIe	Shiga-like toxin IIe (verotoxin 2e, VT-2e)
TS	thymidylate synthase

Terms

IC$_{50}$	concentration of ligand that causes 50% inhibition
K$_i$	dissociation constant for a protein-ligand complex
probe	potential ligand in a docking simulation
RMS	root-mean-square
target	target protein in a docking simulation
ϕ	for carbohydrates the dihedral H1-C1-O1-Cx
ψ	for carbohydrates the dihedral C1-O1-Cx-Hx

Software

AMBER	Assisted Model Building with Energy Refinement (used in Chapters 4 and 5).
BOXSEARCH	A multiple start Monte Carlo docking method (used in Chapters 2, 3, and 5).
DISCOVER	Energy minimization and molecular dynamics program (used in Chapters 2-5).
GEGOP	A conformational search and energy minimization program for carbohydrates and glycopeptides (used in Chapter 4).
GRAPH	A program for scatter (and other) plots (used in Chapters 2 and 3).
INSIGHTII	A program for general visualization and manipulation of molecules. (used in Chapters 2-5).
MOLSCRIPT	A program for molecular diagrams (used in Chapters 4 and 5).
NETWORK	A program for the (re)positioning of mobile polar hydrogens (used in Chapters 2-5).
PROCHECK	A protein structure analysis program (used in Chapter 4).
VADAR	A protein structure analysis program, used here strictly for surface area calculations (in Chapters 2 and 3).

Chapter 1

Introduction

1.1 Protein-ligand interactions

Specific binding interactions between target proteins and their ligands are central to many metabolic processes. Changes in, or disturbances of, particular protein-ligand interactions provide the foundation for many diseases. For example, the disease emphysema is directly related to the interaction between elastase and its endogenous protein inhibitor α_1-antitrypsin [reviewed in (Janoff, 1985)]. In healthy human lungs, a high plasma concentration of α_1-antitrypsin ensures that degradation of lung elastin by released human leukocyte elastase is strictly regulated. However, a genetically-based deficiency of α_1-antitrypsin, or the chemical inactivation of this inhibitor (*eg.* due to smoking), can lead to relatively uncontrolled elastinolysis in the lungs. Over a long period of time this continued process leads to enlargement and loss of elasticity of the lungs - emphysema. In this example the normal process is tightly regulated elastinolysis. An imbalance in the enzyme/inhibitor levels, or a disturbance of the specific interaction by chemical modification of the inhibitor, can, over time, lead to the development of a life-threatening condition. Many other essential, finely balanced, and easily disturbed, metabolic pathways that involve crucial protein-ligand binding

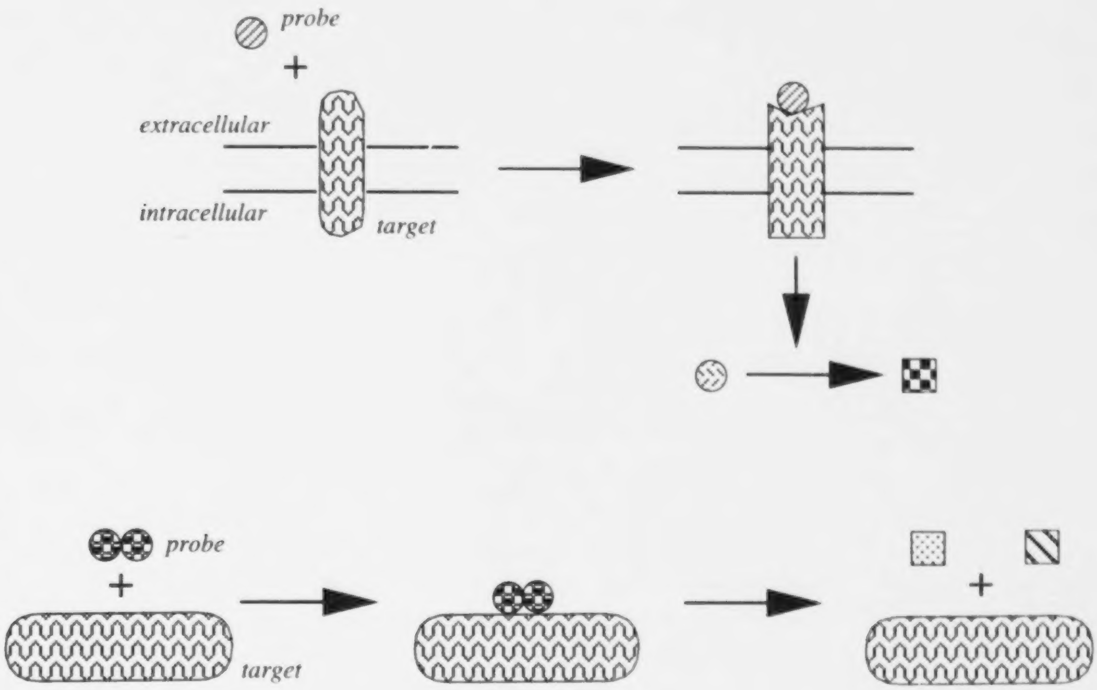


Figure 1.1: Interactions between proteins and their ligands. This schematic shows generalizations of two types of specific interactions. The top figure shows a generalized scheme for the interaction between an extracellular signalling molecule (probe) and its receptor (target). Binding of the signal molecule to its receptor generates an intracellular signal that brings about some downstream effect. The bottom figure shows the binding of a substrate molecule (probe) to an enzyme (target), followed by catalytic cleavage to two different products.

steps are known. The desire to understand how such finely controlled processes work, both normally and in disease states, is one important motivation for studying protein-ligand interactions.

Here, the term "protein-ligand interactions" is used to refer to *specific binding interactions* between a *target* protein and a *probe* or *ligand* molecule. Possible ligands include another protein, peptide, simple or complex carbohydrate, nucleic acid or nucleotide, or an organic small molecule not described by any of the preceding classes (for example, aspirin and acetylcholine). In the elastase/α_1-antitrypsin example discussed above, elastase would be the target and α_1-antitrypsin the ligand. Functionally these interactions include enzyme-substrate, enzyme-inhibitor, and receptor-ligand pairings (Figure 1.1). For the *simulation* of protein-ligand interactions we refer to the ligand as the *probe*, since in many cases we are "probing" for binding sites with molecules or functional groups that may or may not be ligands for that particular target protein.

Another significant motivation for the study of protein-ligand interactions is the field of drug design. Here, the more specific term *ligand design* is preferred, to denote a focus on the specific binding of a ligand or probe to a target protein. Specific ligands for some proteins may be useful as drugs, but here we are concerned only with this first step of *structure-based drug design*. The many subsequent aspects of drug design are not considered here, although the techniques discussed may be applicable to other aspects of this complicated process. Ligand design provides a clear and useful framework within which to discuss modeling and simulation of biomolecules, and we refer to this paradigm for various examples throughout the following discussion. For example, regarding the disease state mentioned above, emphysema, it is thought that specific elastase inhibitors, administered orally or as aerosols, may be useful in halting the progression of the disease. Many different elastase inhibitors, ranging from penicillin analogs to leech proteins, have been investigated for this purpose [*eg.* (Hlasta & Pagani, 1994), and refs therein].

Ligand design is also the subject, to some extent, of Chapters 4 and 5 of this dissertation. In Chapter 4, for example, we focus on certain aspects of a project that involves the investigation of the binding interaction between members of a class of bacterial toxins and their host cell receptors. The toxic effect produced by these bacterial proteins involves initial binding to specific host cell-surface glycolipids, followed by the intracellular release of a catalytic subunit. The action of the the catalytic subunit leads to a disturbance in some aspect of cellular metabolism.

Diseases resulting from such processes include diphtheria, whooping cough, and heomorrhagic colitis (*eg.* hamburger disease). Several of these related toxins and their specific polysaccharide receptors have been structurally characterized [reviewed in (Read & Stein, 1993; Merritt & Hol, 1995)]. Other studies established that inert particles coated with these sugars could bind and inactivate some toxins in cytotoxicity tests [*eg.* (Armstrong et al., 1991)]. These or similar agents may eventually be developed into diagnostic tools and/or therapeutic agents, by virtue of their ability to interfere with the binding interaction of the toxins with their receptors. Further structural and modeling studies (refs in Chapter 4) may aid in the development of stronger and more specific ligands for these toxins.

Classical methods for studying the nature of protein-ligand interactions include the investigation of binding and reaction kinetics, absorption and fluorescence spectroscopy, calorimetry, and other physico-chemical measurements. These techniques allow us to gain an understanding of the various forces that drive and regulate intermolecular binding interactions. In some cases information about a particular residue, or even atom, may be obtained. For example, the influence of ionic strength and *p*H on a reaction rate may provide information about an electrostatic interaction important for binding, or the optimal protonation state for one or more catalytic residues. Absorption and fluorescence experiments can indicate the involvement of specific sidechains in an intermolecular contact. Such methods typically do not allow the fine dissection of a bimolecular complex to the level of specific intermolecular atom-atom contacts. The site-specific mutations facilitated by molecular biology techniques go far in alleviating this limitation, in that changes in phenomena measured by gross physico-chemical techniques can often be ascribed to one or a few side-chain atoms. However, a comprehensive general understanding of protein-ligand interactions requires, in addition to that which can be gained from these more indirect methods, detailed knowledge of a diversity of such interactions at the atomic level.

X-ray crystallography and NMR spectroscopy provide direct structural information about proteins and protein-ligand complexes at the atomic level. Precise measurements of hydrogen-bonds and other electrostatic interactions, as well as hydrophobic contacts, can be obtained. A series of 3-dimensional structures of the same target protein with different ligands can serve as a set of discrete "snapshots", and this may provide information about the range of motions involved in the binding interaction (Shoichet et al., 1993; Greer et al., 1994). Such a series of structures may also provide an explanation for differences in ligand binding affinity and selectivity [several examples are summarized in (Greer et al., 1994)]. This type of information is useful in designing specific ligands for the target protein being studied, and may also be applicable to related targets.

With the aim of gaining a more general understanding of the forces important in protein folding and binding, workers have analyzed large databases of protein structures and compiled statistics describing various aspects of buried and exposed protein surfaces, and protein-protein interfaces (Chothia, 1974; Chothia & Janin, 1975; Miller et al., 1987; Miller, 1989; Argos, 1988; Janin et al., 1988; Korn & Burnett, 1991). These empirical studies have led to *qualitative* generalizations regarding the atom-atom interactions important to protein folding and protein-ligand binding. For example, most protein-protein interfaces are more hydrophobic than the average protein surface, but less so than the average protein interior (Argos, 1988; Janin et al., 1988; Korn & Burnett, 1991). Although the subject is not yet settled, many workers in the field agree that hydrophobic interactions provide a major fraction of the binding energy in biomolecular associations (Kauzmann, 1959; Chothia & Janin, 1975; Dill, 1990; Pace, 1992; Ben-Naim & Mazo, 1993; Creighton, 1993; Rose & Wolfenden, 1993). In this view, van der Waals and electrostatic interactions also make significant energetic contributions, as well as provide for intermolecular specificity by virtue of their precise 3-dimensional arrangement (Chothia & Janin, 1975; Fersht, 1984; Blaney & Dixon, 1993).

A general, widely applicable, *quantitative* description of the forces that control protein folding and binding has not yet been developed. This has been an area of intense research for some time, however, and much progress has been made toward this end [reviewed in (Halgren, 1995; Sippl, 1995)]. One important result of the development of such a general description would be the ability to reliably predict the sites and strengths of protein-ligand associations. This objective is a significant motivator for research and development in this area. Toward this goal, experimental and theoretical studies of biomolecular structure are complementary. The development of theoretical models of protein structure and protein-ligand interaction is highly dependent upon the availability of accurate, experimentally-determined structural information. As theoretical methods emerge and evolve, their ability to reproduce and predict known (experimentally determined) structural features of proteins and protein-ligand complexes can be evaluated. For example, a recent methodological development is consistent with the view (discussed above) that hydrophobic contacts are particularly important in protein-ligand binding. Vakser & Aflalo (1994) adapted their geometric protein-protein docking method to use reduced representations of proteins, eliminating most polar or charged atoms. In docking studies with four test systems, the remaining hydrophobic "partial proteins" gave similar or better results than the full molecular representations.

This dissertation describes studies involving the application of some computer-based simulation approaches to the prediction of biomolecular interactions. Much of the work can loosely be described as "methods testing". This involves using developed tools to answer questions to which the answers are already known. One addresses a problem with the *tool(s)* of interest, and assesses whether or not the correct answer would be accessible in the absence of the (already) known answer. From such work we gain a sense of the limitations of our methods, and insights into how to overcome these limitations. This is essential to the development of reliable simulation methods. We can also, of course, determine what types of problems we have a reasonable chance of solving with the simulation tools currently available. Two of the studies presented here are truly predictive (Chapters 4 and 5), in that we offer predictions of the structures of protein-ligand complexes.

1.2 Simulating protein-ligand interactions

Simulation of protein-ligand binding interactions is a general description that applies to a broad area of research. The computational methods employed in these simulations range from simple to complex. A simple study might involve superposition of one molecule onto the experimentally determined structure of a similar ligand bound to a protein target, followed by manual inspection of the resultant complex. One of the more challenging theoretical methods is free energy perturbation. This method involves the computational transformation of an atom or functional group (*eg.* $H \rightarrow CH_3$, $OH \rightarrow NH_2$), thus providing a method for studying the effects of ligand modification on binding affinity, as well as a variety of other phenomena. These methods have recently been reviewed (Cohen et al., 1990; Cherfils & Janin, 1993; Blaney & Dixon, 1993; Kuntz et al., 1994; Lybrand, 1995). The present discussion is restricted to protein-ligand docking methods, particularly those that are computer-based and automated.

The *docking problem* is commonly described as the prediction of whether or not a probe molecule will bind to a target protein, and, if it does bind, what is the binding mode and what is the affinity of the probe for the target (Wodak & Janin, 1978; Kuntz et al., 1982; Cherfils & Janin, 1993;

Blaney & Dixon, 1993; Hart & Read, 1994; Kuntz et al., 1994). In its simplest form the probe and target molecules are treated as rigid bodies and the docking problem has six degrees of freedom: three translational and three rotational degrees of freedom of the probe relative to the target. In the absence of more than one strict distance constraint, even this simplified problem can be very difficult to solve manually (see below). Manual docking involves a person sitting at a graphics workstation and visually manipulating the probe and/or target molecule. The process is guided by any information available regarding the complex being studied, visual cues, intuition, and whatever intermolecular scoring schemes (see below) are available during the procedure. Automated computational methods are often applied to solving the docking problem (Figure 1.2). At the present time rigid-body docking of the two halves of a protein-protein complex is a widely-studied problem. Different configurations of the complex are generated (the *search*), and the favorability of the different configurations are evaluated (the *score* function). These are the two fundamental aspects of docking. Many automated methods have been developed to address the docking problem [reviewed in (Cohen et al., 1990; Cherfils & Janin, 1993; Blaney & Dixon, 1993; Kuntz et al., 1994; Hart & Read, 1994; Lybrand, 1995)]. The successful solution of many specific docking problems involves the use of both automated and manual docking methods.

A variety of different search methods have been devised for automated docking. These range from relatively biased methods involving very localized searches (*eg.* an active site), to random methods in which all possible orientations of the probe relative to the complete target surface are searched. The Monte Carlo method has been used as a random search method in a variety of multiple minima search problems, including conformational searching, protein folding and automated docking. In Monte Carlo-based rigid-body docking, following equilibration, the energy of the system is evaluated, a random change is applied to the binding mode, and the energy of the

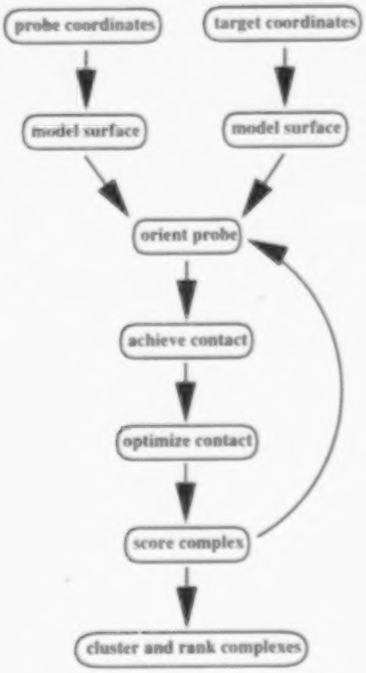

probe coordinates
target coordinates
model surface
model surface
orient probe
achieve contact
optimize contact
score complex
cluster and rank complexes

Figure 1.2: Flowchart of generalized docking algorithm. Many automated docking methods are based upon some type of representation of the surface of one or both of the molecules being docked. Following this, a cycle is commenced, in which an initial configuration of the complex is generated (this may be based on the surface representation) and then optimized. The optimized complex is stored, and the cycle is repeated many times. When the simulation is completed, all of the stored dockings are grouped into clusters (typically based on distance criteria) and then ranked according to whatever score function is used [adapted from (Cherfils & Janin, 1993)].

system is re-evaluated. If the change leads to a decrease in energy, then the new configuration is accepted as the new state of the system, and the cycle is repeated. If the change causes an increase in energy, then a random choice is made to accept or reject the new state of the system. If the energy increase is small relative to the thermal energy of the system then the new state is more likely to be accepted. With extensive sampling, this procedure yields a Boltzmann weighted distribution of the accessible states of the system. Our own automated rigid-body docking method, described in detail in Appendix A, incorporates a combined Monte Carlo and simulated annealing search procedure. The system starts at high temperature, and is slowly cooled down during repeated, short, Monte Carlo cycles. At higher total system energy significant energy barriers are less prohibitive, and the combined procedure allows more extensive searching of configuration space than the Monte Carlo method alone. This simulated annealing cycle is repeated many times from different, randomly chosen, starting configurations. At the end of each cycle the docking is saved if it falls below a user-specified interaction energy cutoff. The interaction energy of a docking is evaluated according to a simple potential energy function. This aspect of simulated docking (scoring) remains much more challenging than the search.

A potential function (or force field) is a mathematical expression that describes the potential energy of a system as a function of its atomic coordinates. For a specific probe-target pair the favored molecular complex represents the global minimum of the potential energy for that system. Therefore, an accurate potential function would allow the simulated *ab initio* prediction of protein folding and binding. The power of such a tool explains the extensive research in this area.

To evaluate new potential functions, the results of simulations based on these functions are compared to experimental results. For example, in an automated docking procedure, a potential function can be used as the score function during the docking search. By studying complexes of known structure, the potential function can be evaluated on the basis of its ability to rank the correct (known) structure as the most energetically favorable. This is a standard test for automated docking methods. As stated above, however, a general and reliable potential function for biomolecular simulation has not yet been developed. This deficiency is manifested in current automated docking studies. For example, in many cases it is relatively easy to accurately re-construct a known complex from its two halves, but much more difficult to achieve the same level of accuracy using the *apo* structure of the probe and/or target molecules. [A recent study (discussed in more detail in Chapter 6;

see also last paragraph of this section) compares several different automated docking methods when applied to the same system, and discusses this aspect of the docking problem (Strynadka et al., 1996)].

Potential functions are being developed to describe the full range of motions and interactions accessible to proteins (and other molecules). An example of a potential function commonly used in biomolecular modeling is given, in simple form, by equation 1.1 (Brooks et al., 1988).

$$E_{total} = E_{1,2pairs} + E_{angles} + E_{dihedral} + E_{vdW} + E_{elec} \tag{1.1}$$

where

E_{total} = the potential energy of the system,
$E_{1,2pairs}$ = the bonded pair potential,
E_{angles} = the bond angle potential,
$E_{dihedral}$ = the dihedral potential,
E_{vdW} = the van der Waals potential, and
E_{elec} = the electrostatic potential.

The first three terms on the right relate to covalent interactions, and are purely *intra*molecular, whereas the final two terms describe non-bonded interactions, which may be both intra- and intermolecular. In rigid-body docking the molecular conformations are completely fixed and *intra*molecular interactions are assumed to be constant, and are therefore ignored in the energy calculation. The potential energy function therefore simplifies to, for example, the Lennard-Jones plus Coulomb form shown in equation 1.2.

$$E_{total} = E_{vdW} + E_{elec} \tag{1.2}$$

where

$E_{vdW} = AR^{-12} - BR^{-6}$, and
$E_{elec} = \frac{q_1 q_2}{\epsilon R}$.

A, B, q_1 and q_2 are the van der Waals parameters and charges for the respective atoms, R is the distance between atom centers, and ϵ is the dielectric constant. Net partial atomic charges are typically derived from charge distribution analysis, as determined by *ab initio* calculations of ground state charge density. The van der Waals parameters are derived from, and optimized against, a variety of experimental and theoretical data pertaining to small molecules comprising the appropriate atom types. Effects of solvent are approximated by a variety of methods, or are often ignored entirely.

Potential functions for protein folding and binding need not be based on first principles (basic physical laws). Potentials of mean force, derived from statistical analysis of experimental data, as well as a variety of other empirical and heuristic methods are also being developed [reviewed in (Sippl, 1995)]. One very simple example relates to the preceding discussion of the importance of hydrophobic surface burial in protein-protein association. In a study of thirty-eight protein-protein complexes, Young *et al.* (1994) found that, in two-thirds of the cases, there was significant overlap between the most hydrophobic cluster of surface residues and the ligand binding site. For all but one of the other complexes the hydrophobic cluster showing more than 30% overlap with the ligand binding site was ranked second, third, or fourth most hydrophobic for that protein (in one case the cluster was ranked sixth). A rule-based scoring scheme derived from this study might be useful in ranking protein-protein dockings. However, our own docking method involves the former type: we use the Lennard-Jones plus Coulomb functional form described above (equation 1.2).

The use of a simplified potential function (*eg.* equation 1.2) in automated docking procedures is necessitated by the many energy calculations that must be performed. Flexibility, explicit solvent, and all-atom representations are just three examples of important properties or parameters that are commonly ignored, or crudely approximated, to speed up the calculations. Flexibility is frequently approximated by simple deletion or truncation of flexible surface sidechains (see Chapter 2). Scaling down van der Waals radii or allowing some interpenetration of atoms ("soft potentials") are other typical methods. Solvent is most often neglected entirely. Aliphatic hydrogens are usually modeled implicitly by using scaled carbon radii. Polar hydrogens are modeled explicitly, but the extreme mobility of these atoms is often completely ignored in the automatic phase of the docking study. Charge representation is also an area that needs much improvement. Indeed, when one considers all of the approximations made in docking simulations, the accuracy of many of the results reported to date is surprising. It should be stressed that many improvements will be achieved simply by gains in available computing power. For example, more general application of many "brute force" methods (*eg.* systematic conformational search) will become feasible in the next few years. As theoretical understanding, computer power, and algorithmic design all continue to improve, so too will the potential functions used in docking. Theoretical and methodological advances should lead to increased accuracy and greater generality.

Recalling the preceding discussion about the importance of hydrophobic effects in protein-ligand binding, the lack of consideration of solvent effects in the simple potential function used in our docking method is an obvious, possibly critical, limitation of the method. Another is the rigid-body assumption. The incorporation of conformational flexibility into automated docking methods is one of the current frontiers of research in this area, but the research presented in later Chapters does not address this problem directly (a brief discussion is presented in Appendix B). On the other hand, we do report the results of studies aimed at approximating desolvation or hydrophobic effects in automated docking. In relation to this work, one example of a particularly relevant empirical method is briefly described here, and in more detail in Chapter 3.

Eisenberg & Mclachlan (1986) brought together the results of many workers (reviewed in some detail in Chapter 3), and developed a method for evaluating protein folding and binding based on surface burial. An energy value, derived from transfer experiments with small amino acid derivatives, is assigned to the surfaces of different atom types. Energetic penalties and rewards are then calculated based on the total buried and exposed areas of the various surface types. Folded proteins and protein-protein complexes are evaluated on the basis of the sum of these energies, the so-called "solvation free energy". This term has been useful in distinguishing between correctly and incorrectly folded proteins (Eisenberg & McLachlan, 1986; Chiche et al., 1990), but was not found to be generally applicable as a score function in rigid-body protein-protein docking (Shoichet & Kuntz, 1991). More recently, workers have incorporated this or similar terms into standard *in vacuo* potential functions, as a means of implicitly modeling solvent or hydrophobic effects. We tested the effect of this correction on the analysis of docking results obtained with the simple potential function described above (equation 1.2; Chapter 3).

1.3 Some applications of automated docking

The research presented in the following chapters of this dissertation involves the *application* of automated docking, and other molecular modeling techniques, to the solution of biological questions. It is therefore appropriate to complete this brief background with a final section describing a few problems that have been studied with similar techniques, and the methods used. This should provide some feel for the types of problems that can be addressed, the limitations of these methods, and the information that can be obtained.

One of the earliest reports of a computer-based automated docking procedure came from Wodak & Janin (1978), and this method of *protein-protein* docking continues to be refined (Cherfils et al., 1991; Cherfils et al., 1994). Two key approximations were made to facilitate the calculations. First, the models of the proteins used were simplified by representing each residue of the target and the probe as single spheres, or "interaction centres", centred on the centre of mass of the sidechain. Second, no flexibility of the molecules was considered. (It is interesting, and also cautionary, to note that despite the tremendous advances made in available computer power and algorithm design, similar approximations still limit many current docking simulations - see Appendix B for a discussion of some recent algorithmic advances.) A spherical grid surrounding the target was constructed, and probe-target configurations were sampled for each point on the grid, by manipulating the probe at each of the grid points. For each grid point of this systematic search, the docking with the greatest number of interaction centre-interaction centre contacts was energy minimized. The simple potential function used included a van der Waals-like term that allows some interpenetration of the simplified residues, and an approximate desolvation term related to buried surface area (which in turn is approximated from the simplified protein model).

This early study yielded several significant results, three of which are of particular relevance to the present discussion. First, an extremely simple model of protein-protein interaction can reproduce the experimentally observed configuration for a probe-target complex. Although generally not as crude as the "interaction centres" used in this method, current methods use various simplifying models to achieve reasonable computation times. These include little or no consideration of flexibility, united atom representations, no consideration of solvent effects, and grid-based energy calculations. Second, although the crystallographically-observed correct answer was ranked amongst the lowest energy dockings, it was energetically indistinct from several clearly incorrect dockings. While the sampling of configuration space in rigid-body docking would seem to be a solved problem, at least within reasonable limits, the ability to energetically distinguish correct dockings remains a significant challenge.

Third, total buried surface area was a good indicator of complex stability, but, like the simple potential function (which incorporates a term related to this parameter), this parameter was also unable to clearly differentiate the correct answer from several incorrect ones (also discussed in preceding section).

A more recent version of this method (Cherfils et al., 1991) addresses some of the limitations of the original method. The systematic search, which necessitated relatively coarse sampling of the accessible configuration space (Wodak & Janin, 1978), has been replaced by a Monte Carlo approach (the Monte Carlo method was described above, and is also discussed in Appendix A). Presumably this allows finer sampling of the more favorable regions of the search space. Another significant development is the elimination of the crude "interaction centres" protein model from the later stages of the simulation. After the dockings have been generated and divided into clusters of similar dockings (clustering is described in detail in Appendix A), the simplified representatives of each cluster are replaced by full representations of the proteins. These models are then energy minimized with a more sophisticated potential function, with sidechain flexibility allowed for interface residues.

Janin and co-workers have reported successes in docking with several different biological systems (Wodak & Janin, 1978; Cherfils et al., 1991; Cherfils et al., 1994) since the initial development of this method (Wodak & Janin, 1978). Recently they used mutation information to predict a few possible dockings for a hemagglutinin-antibody complex (Cherfils et al., 1994). The structures of each of the isolated components is known, but that of the complex is still under investigation. This prediction provides a "real" test of this method, and automated docking in general.

Two significant possible limitations of this particular method require mention here. First, no applications of this method to the docking of small (synthetic or biological) molecules to a protein target have been reported, and the simple molecular model used in the initial stages of the docking search may restrict application of this method to protein-protein docking. Whether or not this limitation is restrictive is debatable - many people interested in applying automated docking methods are interested in both protein-protein and protein-small molecule docking. Second, in many of the reported applications the search was restricted to a relatively small fraction of the total surface area of one or both of the proteins. In a "real" application of docking the extent to which the search can be confined depends on the distance constraints available (discussed in Chapters 2-6; an ideal docking method would search the total surface of both molecules, and rank the correct complex as the most energetically favorable - this point is discussed in some detail in Chapters 2 and 6).

Probably the most widely used automated docking method is DOCK, which has been under continual development since 1982 by Kuntz and co-workers [*eg.* (Kuntz et al., 1982; DesJarlais et al., 1986; Shoichet et al., 1992; Leach & Kuntz, 1992; Shoichet & Kuntz, 1993)]. This method is useful for both protein-protein and protein-small molecule docking. DOCK starts from a description of the shape of the target molecule. The solvent accessible surface (Richards, 1977) of the target is calculated (Connolly, 1983), and possible (or known) binding sites are described as sets of spheres that overlap each other and occupyconcave regions of the protein surface (Kuntz et al., 1982). The ligand is then represented as a set of spheres centred on its component atoms, and a distance matching procedure is used to find ligand orientations that match the representation of the active site. DOCK has undergone considerable evolution since its initial development (see citations above). The original method employed a completely rigid target and probe(s), and a relatively crude scoring function with overlap and hydrogen-bonding terms (Kuntz et al., 1982). Flexibility was then modeled by breaking down test ligands into rigid fragments, docking the fragments, and then rejoining them at a later stage of the docking simulation (DesJarlais et al., 1986). An exciting development was the application of the DOCK method to screening of a database of small molecule structures as a method for finding novel ligands for target proteins [(DesJarlais et al., 1988); see next paragraph]. Over the course of these developments the practice of employing the AMBER modeling package (Weiner et al., 1984) for energy refinement of favored complexes discovered by the cruder evaluation methods was adopted (DesJarlais et al., 1986; DesJarlais et al., 1988). In 1991, Shoichet & Kuntz produced a comprehensive and instructive report on protein-protein docking that described the evaluation of several scoring functions including similarity to the crystal structure, buried surface area, surface area-based solvation free energy, packing, mechanistic filters, electrostatic interaction energy, and molecular mechanics. They concluded that while simpler methods were adequate in some cases, molecular mechanics was the only scoring method that consistently ranked the correct answer amongst the most favorable. However, even this relatively sophisticated approach could not reliably distinguish between low energy correct and incorrect complexes. A comparable test (to that above for protein-protein docking) of various score functions has not been reported for systems involving small molecule docking to protein targets. More recent versions of DOCK have employed improved shape-matching procedures (Shoichet et al., 1992), chemical complementarity (Shoichet & Kuntz, 1993), grid-based energy evaluation (Meng et al., 1992), and conformational flexibility of the ligand (Leach & Kuntz, 1992).

A recent report (Shoichet et al., 1993) described the use of the DOCK program as a tool for searching a database of small molecule structures for inhibitors of thymidylate synthase (TS). This enzyme is a therapeutic target for proliferative diseases, including cancer. Ligand ranking in this study involved an initial measure of steric fit, calculation of the electrostatic interaction energy, and, finally, application of a solvation correction for the higher ranking “hits”. The initial round of searching ranked some known inhibitors of TS favorably, and also identified what turned out to be several novel TS inhibitors. Results of this initial round of searching showed very poor correlation between calculated and measured affinity *for structurally diverse compounds* [Table 1 in (Shoichet et al., 1993)]. Whether or not this statement applies to structurally and chemically similar ligands is not clear. The structure of TS complexed with one of the novel inhibitors, sulisobenzone (SB), was determined crystallographically under two sets of conditions. It was found that the observed binding mode was affected quite dramatically by the choice of crystallization buffer, and that under both conditions the binding mode differed from that predicted by the docking simulation. This suggested exploration of a previously neglected region of the binding pocket. A similarity search for SB-like compounds, followed by two progressive cycles of DOCK-based database searching, led to the identification of several phenolpthalein (PTH) derivatives as novel TS ligands. These compounds had TS IC_{50} values as low as 3 μM (PTH had an IC_{50} of 15 μM and a K_i of 4.4 μM). Crystallographic solution of the TS-PTH complex showed much better agreement between the calculated and observed PTH binding modes than that noted for the SB-TS complex. The disparity between certain calculated and experimental aspects of this study are less than satisfying. However, novel inhibitors with K_i values near 1 μM were identified amongst commercially available compounds. This represents a significant achievement for computer-based methods of ligand discovery. Also, the combined crystallographic and simulation results suggest several modifications that might lead to better binding. This type of rational modification in the early stages of lead optimization, as well as the ability to perform such database searches, on either proprietary or commercially accessible databases, represents an attractive alternative, or complement, to the more traditional methods employed in this aspect of drug discovery.

In a completely different application the DOCK program was used to model the binding modes of several series of structurally-related acetylcholinesterase (AChE) inhibitors (Yamamoto et al., 1994). The goals of this work were to understand the observed structure-activity relationships (SARs) for these classes of inhibitors, and also (presumably) to suggest structure-based modifications that might lead to the development of more potent inhibitors. A modified version of DOCK, known as directed-DOCK (Leach & Kuntz, 1992), that allows systematic conformational searching of part of the ligand molecule while another part of the molecule is fixed, was employed in this study. Docking simulations with partly flexible/partly fixed ligands bound at the active site of AChE were used to generate many possible AChE-ligand complexes. The more energetically-favorable of these were then energy minimized. There were many limitations to this work - chief among these were the use of a completely rigid enzyme at all times, and the lack of experimental structures of any of the ligands used (or a closely related analog; however, one of the compounds does act as a label for the active site Ser of AChE, thus imposing a certain intermolecular distance constraint on the interaction). Despite these significant limitations, the authors were able to derive model complexes for several series of compounds that were consistent with the observed SARs for these compounds. While this consistency does not prove the accuracy of the modeling results, it may be useful in further design of AChE inhibitors.

Prediction of a biologically important protein-protein complex, given only the structures of the complex components, is an important application of computational docking simulations. Stoddard & Koshland (1992) used a Monte Carlo-based docking method [(Goodsell & Olson, 1990); see also Appendix B] and information derived from mutational studies to predict the structure of the maltose binding protein (MBP) bound to the aspartate receptor from *E. coli* (a detailed description of the Monte Carlo method is provided in Appendix A, where our own docking method is outlined). Certain residues of MBP had previously been shown, by mutational analysis, to be essential to the interaction of MBP with the aspartate receptor [summarized in (Stoddard & Koshland, 1992)]. Two octapeptides containing these key residues were excised from the MBP structure (docking the whole MBP probe was considered impractical with the method used), and docked to the ligand-binding domain of the receptor. The entire target protein and the backbones of the probes were rigid; flexibility was allowed only in the probe sidechains. For each octapeptide probe one solution dominated the docking results. These solutions were shown to be consistent with docking of the complete MBP structure to the receptor, and a final model of the complex was generated by superimposing the whole protein onto the docked octapeptides. This generated two minor steric clashes involving surface loops of the receptor (the receptor used was a model developed from the structure of the 80% identical *Salmonella* receptor) that were readily alleviated by restrained energy minimization. The final model was consistent with the structural inferences derived from a variety of mutational studies. The experimental characterization of this complex remains to be done. This study provides a good example of how biochemical information can be combined with limited structural information to reduce a difficult docking problem to a more manageable level.

Docking may also serve as a tool to aid in structural refinement. Goodsell et al. (1993) docked flexible ligands to the active site of aconitase to generate models of bound citrate and isocitrate (using the method referred to in the preceding paragraph). These models were then used in the structural refinement of the complexes. One of the four isocitrate conformers generated was found to "unambiguously fit the observed density". Two citrate conformers were generated, and one of these was found to be similar to that of nitrocitrate (a citrate isostere) bound to aconitase. Both conformers differed from that of a previous model, which was generated using the structure of the native (uncomplexed) enzyme. In this study, several models of the catalytic intermediate *cis*-aconitate bound at the aconitase active site were studied, and two of these seemed likely based upon their favorable interaction energies (relative to the other models), and also their similarity to bound citrate and isocitrate.

A comprehensive survey of docking methods and applications is well beyond the scope of this dissertation. In addition to those methods described above, some more recent developments are presented in Appendix B, in the context of a discussion concerning docking methods that consider molecular flexibility.

Finally, the results of a protein-protein "docking challenge" issued to the "docking community" were recently published [(Strynadka et al., 1996); following completion of most of this dissertation], and are particularly relevant to this dissertation. The results obtained with different docking methods when applied to the same problem can be directly compared; for most docking studies this is not possible. A summary and discussion of this report (Strynadka et al., 1996) is presented in Chapter 6, and compared to some of the results presented in this dissertation.

1.4 Summary of the Introduction

The importance of protein-ligand binding interactions has been discussed, with particular reference to disease conditions. The *docking problem* was defined, and some computer-based approaches to solving the problem were presented. A brief description of the automated docking method used in some of the research presented in later chapters of this dissertation was given. Varieties of the potential energy functions used in biomolecular modeling and docking methods were summarized, and the limitations of these functions were stressed. Examples of recent methodological developments were used throughout, in an effort to relate the discussion to current practice. Finally, several recent applications of computer-based docking to problems of structural biology and ligand design were summarized, to again stress the practical uses of the method in general.

My own research has centred around the application of molecular modeling and docking methods to several different problems. Chapter 2 presents the results of docking simulations with variants of the diubiquitin system. This chapter includes a detailed discussion of methods of analysis of protein-protein docking results. In the final study of this project we analyzed the hydrophobicities of the dimer interfaces, as an aid to distinguishing correct from incorrect dockings. This led to a more detailed investigation of the use of surface burial as a method of approximating desolvation and hydrophobic effects in protein-ligand binding, presented in Chapter 3. In Chapter 4 we model the binding interaction between a bacterial toxin and its specific carbohydrate ligand, and offer an explanation for the change in binding specificity conferred upon the lectin by a double mutation. Finally, in Chapter 5, a novel application of fragment-based docking is used to contruct a model of the complex of NAD bound to pertussis toxin, based on the related structure with diphtheria toxin. Similarities and differences between the two complexes are noted and discussed.

Two appendices are included, to extend the discussion of two topics covered briefly in this Introduction. Appendix A contains a detailed description of both the theory and practice of the BOXSEARCH docking program. Appendix B briefly summarizes and surveys methods that have been used to simulate conformational flexibility in automated docking studies.

1.5 References

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Prot. Eng.*, 2:101–113.

Armstrong, G. D., Fodor, E., & Vanmaele, R. (1991). Investigation of shiga-like toxin binding to chemically synthesized oligosaccharide sequences. *J. Infect. Dis.*, 164:1160–1167.

Ben-Naim, A. & Mazo, R. M. (1993). Size dependence of the solvation free energies of large solutes. *J. Phys. Chem.*, 97:10829–10834.

Blaney, J. M. & Dixon, J. S. (1993). A good ligand is hard to find: automated docking methods. *Perspect. Drug Discov. Des.*, 1:301–319.

Brooks, C. L. I., Karplus, M., & Pettitt, B. (1988). Proteins: a theoretical perspective of dynamics, structure, and thermodynamics. *Adv. Chem. Phys.*, 71:23–30.

Cherfils, J., Bizebard, T., Knossow, M., & Janin, J. (1994). Rigid-body docking with mutant constraints of influenza hemagglutinin with antibody HC19. *Proteins: Structure, Function, and Genetics*, 18:8–18.

Cherfils, J., Duquerroy, S., & Janin, J. (1991). Protein-protein recognition analyzed by docking simulation. *Proteins: Structure, Function, and Genetics*, 11:271–280.

Cherfils, J. & Janin, J. (1993). Protein docking algorithms: simulating molecular recognition. *Curr. Op. Struc. Biol.*, 3:265–269.

Chiche, L., Gregoret, L. M., Cohen, F. E., & Kollman, P. A. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl. Acad. Sci. U.S.A.*, 87:3240–3243.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248:338–339.

Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature*, 256:705–708.

Cohen, N. C., Blaney, J. M., Humblet, C., Gund, P., & Barry, D. C. (1990). Molecular modeling software and methods for medicinal chemistry. *J. Med. Chem.*, 33:883–894.

Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713.

Creighton, T. E. (1993). *Proteins: structures and molecular properties*. W. H Freeman and Company, New York, 2 edition.

DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D., & Venkataraghavan, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem*, 29:2149–2153.

DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D., & Venkataraghavan, R. (1988). Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.*, 31:722–729.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29:7133–7155.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, 319:199–203.

Fersht, A. R. (1984). Basis of biological specificity. *Rends Biochem. Sci.*, 9:145–147.

Goodsell, D. S. & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct, Funct, Genet*, 8:195–202.

Greer, J., Erickson, J. W., Baldwin, J. J., & Varney, M. D. (1994). Application of the three-dimensional structures of protein target molecules in structure-based drug design. *J. Med. Chem.*, 37:1035–1054.

Halgren, T. A. (1995). Potential energy functions. *Curr. Op. Struc. Biol.*, 5:205–210.

Hart, T. N. & Read, R. J. (1994). Multiple-start Monte Carlo docking of flexible ligands. In Merz, K. M., J. & LeGrand, S. M., editors, *The Protein Folding Problem and Tertiary Structure Prediction*, pages 71–108. Birkhäuser, Boston.

Hlasta, D. J. & Pagani, E. D. (1994). Human leukocyte elastase inhibitors. *Ann. Rep. Med. Chem.*, 29:195–204.

Janin, J., Miller, S., & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.*, 204:155–164.

Janoff, A. (1985). Elastases and emphysema. Current assessment of the protease-antiprotease hypothesis. *Am. Rev. Resp. Dis.*, 132:417–433.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Prot. Chem.*, 14:1–63.

Korn, A. P. & Burnett, R. M. (1991). Distribution and complementarity of hydropathy in multisubunit proteins. *Proteins: Struct. Funct. Genet.*, 9:37–55.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288.

Kuntz, I. D., Meng, E. C., & Shoichet, B. K. (1994). Structure-based molecular design. *Acc. Chem. Res.*, 27:117–123.

Leach, A. R. & Kuntz, I. D. (1992). Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comp. Chem.*, 13:730–748.

Lybrand, T. (1995). Ligand-protein docking and rational drug design. *Curr. Op. Struc. Biol.*, 5:224–228.

Meng, E. C., Shoichet, B. K., & Kuntz, I. D. (1992). Automated docking with grid-based energy evaluation. *J. Comp. Chem.*, 13:505–524.

Merritt, E. A. & Hol, W. G. J. (1995). AB_5 toxins. *Curr. Op. Struc. Biol.*, 5:165–171.

Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins. *Prot. Eng.*, 3:77–83.

Miller, S., Janin, J., Lesk, A., & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.*, 196:641–656.

Pace, C. N. (1992). Contribution of the hydrophobic effect to globular protein stability. *J. Mol. Biol.*, 226:29–35.

Read, R. J. & Stein, P. E. (1993). Toxins. *Curr. Op. Struc. Biol.*, 3:853–860.

Richards, F. M. (1977). Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.*, 6:151–176.

Rose, G. D. & Wolfenden, R. (1993). Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu. Rev. Biophys. Biomol. Struct.*, 22:381–415.

Shoichet, B. K., Bodian, D. L., & Kuntz, I. D. (1992). Molecular docking using shape descriptors. *J. Comp. Chem.*, 13:380–397.

Shoichet, B. K. & Kuntz, I. D. (1991). Protein docking and complementarity. *J. Mol. Biol.*, 221:327–346.

Shoichet, B. K. & Kuntz, I. D. (1993). Matching chemistry and shape in molecular docking. *Prot. Eng.*, 6:723–732.

Shoichet, B. K., Stroud, R. M., Santi, D. V., Kuntz, I. D., & Perry, K. M. (1993). Structure-based discovery of inhibitors of thymidylate synthase. *Science*, 259:1445–1450.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Op. Struc. Biol.*, 5:229–235.

Stoddard, B. L. & Koshland, D. E. J. (1992). Prediction of the structure of a receptor-protein complex using a binary docking method. *Nature*, 358:774–776.

Strynadka, N. C. J., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R. amd Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M., & James, M. N. G. (1996). Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. *Nature Struct. Biol.*, 3:233–239.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. J., & Weiner, P. (1984). A new force-field program for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784.

Wodak, S. J. & Janin, J. (1978). Computer analysis of protein-protein interactions. *J. Mol. Biol.*, 124:323–342.

Yamamoto, Y., Ishihara, Y., & Kuntz, I. D. (1994). Docking analysis of a series of benzylamino acetylcholinesterase inhibitors with a phthalimide, benzoyl, or indanone moiety. *J. Med. Chem.*, 37:3141–3153.

Chapter 2

Monte Carlo docking with ubiquitin[1]

2.1 Introduction

A variety of computer-based methods for the simulation of bio-molecular docking has been reported (for reviews see Cherfils & Janin, 1993 or Kuntz et al., 1994) and this is currently an active area of research for many groups, including our own. It seems reasonable to say, however, that the development of such methods as tools for the solution of real biological problems is just beginning. The development of more accurate and robust docking algorithms requires the study of a diverse selection of biological systems, as well as critical examination of the effects of various approximations used during the simulations. The ubiquitin conjugation system provides an opportunity to study a variety of protein-protein interactions and in the present work we report the results of docking simulations, using an algorithm under development in this laboratory, with the ubiquitin/diubiquitin system.

---

[1] A version of this chapter has been published: M.D. Cummings, T.N. Hart, & R.J. Read, 1995, *Monte Carlo docking with ubiquitin*, Protein Sci. *4*:885-899. Reprinted with the permission of Cambridge University Press.

The regulated degradation of specific proteins is one of the fundamental processes that enable cells to change rapidly from one metabolic state to another. Covalent attachment of ubiquitin polymers to protein substrates appears to be one of the major pathways by which cellular proteins are preferentially targeted for degradation in eukaryotic cells (for reviews see Hershko & Ciechanover, 1992; Hochstrasser, 1992; Jentsch, 1992; Varshavsky, 1992). Since ubiquitin is conjugated to a variety of protein substrates it seems reasonable that this selectivity is not a function solely of ubiquitin but derives, at least in part, from features of the enzymes involved in ubiquitin conjugation and/or features of the substrate proteins. This is supported by the existence of a large family of ubiquitin-conjugating enzymes (E2s; Hochstrasser, 1992; Rechsteiner, 1991) and the possibility of a similarly large family of ubiquitin-protein ligases (E3s; Rechsteiner, 1991). Rechsteiner (1991) has postulated that ubiquitin may act as a "movable binding site", thus facilitating the interaction, or at least spatial proximity, of proteins which are not complementary to each other.

Ubiquitin is a highly conserved protein found in all eukaryotic cells. The minor sequence variations of plant and yeast ubiquitin are confined to one region of the protein (Vijay-Kumar et al., 1987) and, consequently, it was suggested that this part of the protein surface is not involved in recognition events during conjugation and/or proteolysis (Wilkinson, 1988). Subsequently it was shown that this region was indeed distant from the dimer interface in diubiquitin (Cook et al., 1992a). Chemical modification studies have indicated that several residues may be crucial to the interactions involved in activation, conjugation, or proteolysis (Wilkinson, 1988). The information available is, at best, suggestive of the relative importance or unimportance of certain regions of the ubiquitin molecule in the various intermolecular interactions that are involved in ubiquitin-dependent proteolysis. An understanding of the features involved in the interaction of ubiquitin with the various enzymes of the ubiquitin conjugation pathway would help to explain some of the differences observed among the various forms of these enzymes (see, for example, Hochstrasser, 1992; Hershko & Ciechanover, 1992).

Ubiquitin activating enzymes (E1s) and E2s form thioester linkages between a catalytic cysteine residue and the carboxy-terminal glycine of ubiquitin. The catalytic cysteine has been identified as Cys 88 in the E2 UBC1 isolated from the plant *Arabidopsis thaliana* and is located in a region that shows a relatively high degree of sequence conservation among E2s from several sources (Cook et al., 1992b). The crystal structure of this (Cook et al., 1992a) and another E2 (Cook et al., 1993) have recently been reported. Ubiquitination sites have been mapped to specific residues or regions of two degradation target proteins for which structures have been determined (Sokolik & Cohen, 1992; Hill et al., 1993). Crystal structures have also been reported for ubiquitin and Gly A76 - Lys B48 isopeptide-linked diubiquitin (entries 1ubq and 1aar, respectively, in the Brookhaven Protein Data Bank (Bernstein et al., 1977); A and B refer to the two distinct ubiquitin monomers in diubiquitin). During the course of this work the structure of tetraubiquitin was also reported (Cook et al., 1994).

Evidence indicates that the Gly A76 - Lys B48 isopeptide bond is the linkage of major importance in the ubiquitin polymers that target substrate proteins for degradation (Chau et al., 1989; Gregori et al., 1990), and this is the only linkage observed in the diubiquitin and tetraubiquitin structures. The observed twofold pseudosymmetry of the diubiquitin structure does not allow for further extension to higher polymers; however, the tetraubiquitin structure can be extended indefinitely. The flexibility of the C-terminus of the ubiquitin molecule allows a pair of covalently linked monomers access to a variety of configurations.

The biological relevance of the various polymeric states of ubiquitin is unclear. Monoubiquitination can apparently support degradation in some cases (Gregori et al., 1985; Hershko & Heller, 1985), and E2's vary in their ability to transfer ubiquitin polymers to free and ligated (to a target protein) monoubiquitin (Chen & Pickart, 1990). Diubiquitin acts as a steady-state intermediate during synthesis of higher order polymers by an E2 (Chen & Pickart, 1990). A quantitative study of the targetting efficiency of ubiquitin polymers of varying length has not been reported. One of the subunits of the proteolytic complex that degrades ubiquitinated proteins has been shown to bind ubiquitin polymers cooperatively with respect to chain length (Deveraux et al., 1994). While it seems clear that ligation of a relatively large multiubiquitin chain to a protein can target that protein for degradation by the 26S proteasome, the functions and relative importance of the various polymeric forms of ubiquitin are currently unknown. Cook et al. (1994) state that the crystal structure of diubiquitin probably represents the predominant solution structure, and that the polymer likely undergoes a configurational "switch" to the tetraubiquitin-like configuration when a third monomer is conjugated to the growing polymer. We can find no evidence which argues against such conclusions. Although the current work is not directly concerned with clarifying these issues, our own results are consistent with the suppositions of Cook and co-workers (Cook et al., 1994).

Since the present work is concerned with the prediction of biomolecular complexes, it is interesting to note that the observed diubiquitin structure (Cook et al., 1992a) resembles the earlier qualitative prediction of Silver et al. (1992). With the few exceptions noted above there is little direct structural information available regarding the nature of the ubiquitin binding sites on the enzymes involved in ubiquitin conjugation, on the target proteins to which ubiquitin is conjugated, or on the proteases which recognize ubiquitinated proteins. A similar lack of information exists regarding the affinity of ubiquitin for itself or for other proteins.

We have studied the structure of diubiquitin using the 2.3 Å resolution crystal structure of diubiquitin (1aar; Cook et al., 1992), the 2.4 Å structure of tetraubiquitin (1tbe; Cook et al., 1994), and the 1.8 Å structure of ubiquitin (1ubq; Vijay-Kumar et al., 1987). One of our long-term goals is to predict the structure of complexes involved in ubiquitin conjugation. The ubiquitin/diubiquitin system, with which we could test our ability to predict a known answer using the structures of both the complexed and uncomplexed monomers, seemed to be a logical starting point for such studies. Given that ubiquitin is known to interact specifically with numerous apparently non-homologous enzymes, and that the affinity of one ubiquitin molecule for another is quite low (see below), we consider this system to be an especial challenge for docking methods in general. As a bonus, our initial results indicated that the ubiquitin/diubiquitin system would be very instructive for the development and evaluation of docking strategies. Despite the marked structural similarity between the ubiquitin monomer and each of the two halves of the diubiquitin structure we were unable to predict the diubiquitin structure with the unmodified ubiquitin monomer. Truncation of a flexible residue (Arg42) previously implicated as being crucial to one or more aspects of ubiquitin-dependent proteolysis facilitated the prediction of a dimer configuration similar to that of the experimentally observed diubiquitin molecule.

2.2 Materials and methods

2.2.1 Structures

The structures of ubiquitin (1ubq; Vijay-Kumar et al., 1987), diubiquitin (1aar; Cook et al., 1992), and tetraubiquitin (1tbe; Cook et al., 1994) were from the Brookhaven Protein Data Bank (PDB; Bernstein et al., 1977).

2.2.2 Hardware

All calculations were performed on a Silicon Graphics R4000 Crimson or R4000PC Indy.

2.2.3 Software

Docking simulations were performed with the program BOXSEARCH, which is under development in this laboratory [(Hart & Read, 1992); see Appendix A]. Monte Carlo minimizations were performed with a slightly modified version of BOXSEARCH. Various tools for the analysis of docking results have been developed in this laboratory. Systematic conformational searches, energy minimizations, as well as general structure manipulation and visualization were performed with DISCOVER and various modules of the INSIGHTII program (Biosym Technologies, San Diego). Polar hydrogen positions were optimized with the NETWORK program (Bass et al., 1992) prior to energy minimization. Some superimpositions were done according to the method of Rao and Rossmann (Rao & Rossmann, 1973). Surface area calculations were performed with the VADAR program (under development at the University of Alberta; personal communication from D.S. Wishart) which incorporates the ANAREA program (Richmond, 1984). Scatter plots were prepared with the GRAPH module of the program SETOR (Evans, 1993).

2.2.4 Structure preparation

Water molecules were removed and hydrogens were added to the PDB structures according to the standard method in INSIGHTII at neutral pH. Any residue deletions or sidechain truncations were done at this time. Polar hydrogens were then repositioned by the program NETWORK (Bass et al., 1992), which maximizes intramolecular hydrogen bond networks (in this case intramolecular hydrogen bonds were not affected by the deletion of the waters prior to running NETWORK). The polar hydrogen positions were then further optimized by 200 cycles of steepest descents energy minimization followed by a maximum of 200 cycles of conjugate gradient energy minimization with the CVFF forcefield in DISCOVER. Minimizations were done *in vacuo* with a dielectric constant of 1.0, and only hydrogen atoms were allowed to move. In the case of the two halves of the diubiquitin structure each half was treated separately so as to avoid biasing any hydrogen positions in favor of a particular docking. We consider the structure being *docked to* to be the *target* and the structure being *docked onto the target* to be the *probe*. Since our docking protocol does not allow for covalent bonds between the target and the probe we deleted the C-terminal residue (Gly 76) from both the target and the probe in all of our docking experiments.

2.2.5 Reference structures

In all of the present experiments we had a "correct" answer which we sought in our docking simulations. For the reconstruction of diubiquitin we superimposed the two independently prepared halves of the structure onto the experimentally determined diubiquitin structure and then subjected the probe to rigid-body Monte Carlo minimization with the annealing schedule shown in Table 2.1. We performed one set of experiments in which we used a copy of the target as the probe. For this experiment, as well as that involving construction of diubiquitin from two ubiquitin monomers, we followed a procedure identical to that described above. The configurational space within which the docking searches took place was identical for all of the experiments reported here.

step #	kT (kcal/mol)	# runs	max. rotation (degrees)	max. translation (Å)
1	10^{-3}	500	3.0	1.0
2	10^{-4}	1000	1.0	0.2
3	10^{-5}	1000	0.5	0.05

Table 2.1: The Monte Carlo minimization schedule.

2.2.6 Docking

Docking simulations were performed essentially as described (Hart & Read, 1992) with the annealing schedule shown in Table 2.2. Very briefly (a detailed description of the theory and use of this program is given in Appendix A), a docking "run" with BOXSEARCH commences with the random placement of the probe within a search space which includes all or part of the target molecule. Rigid-body Monte Carlo-based simulated annealing is then performed on the probe-target configuration, according to an annealing schedule which specifies a fixed number of Monte Carlo steps at each temperature (Table 2.2). Dockings which fall below a user-specified interaction energy cutoff are written to output. A typical docking experiment consists of several thousand such "runs". All of the present experiments were performed in a 49 Å cube which excluded one "face" of the target and allowed for all possible orientations of the probe relative to a large part of the target surface (see below).

step #	kT (kcal/mol)	# runs	max. rotation (degrees)	max. translation (Å)
1	10	5	18	5.0
2	8.0	5	18	5.0
3	6.0	5	18	5.0
4	4.0	5	18	5.0
5	2.0	5	18	5.0
6	1.0	5	18	5.0
7	0.5	10	18	5.0
8	0.25	10	18	5.0
9	0.1	50	9	2.5
10	10^{-4}	50	9	2.5

Table 2.2: The annealing schedule for Monte Carlo docking. Dockings that pass the energy cutoff after this Monte Carlo run repeat step 10 four times.

2.2.7 Analysis - reconstruction of diubiquitin from its two halves

All dockings were compared to the appropriate reference structure on the basis of energy and all-atom RMS differences. Cluster analysis was used to group the dockings into clusters or families. We saved the lowest energy member of each family and counted the number of dockings that were in each family (within 2 Å RMS of the lowest energy family member).

2.2.8 Analysis - construction of diubiquitin from two ubiquitin monomers

In addition to the analyses described in the previous section we applied several more critical data filters to the results obtained in these experiments. The rotations and translations necessary to superimpose dockings onto the appropriate reference structure were determined as a complement to the more straightforward, but at times less informative, RMS differences (Shoichet & Kuntz, 1991). The rotation necessary to superimpose a docking onto the target was also determined as a measure of the pseudo-two-fold symmetry of the dockings. Similar to Shoichet and Kuntz (1991) who employed mechanistic filtering to rule out incorrect dockings, we used the PROBE:75:C to TARGET:48:NZ distance to rule out certain configurations, based on the presumed difficulty of forming the necessary isopeptide bond between distant atoms. In one case a systematic conformational search was carried out on the Lys 48 sidechain of the target as well as the flexible C-terminus of the probe. Changes in exposed surface area upon complex formation were calculated for some experiments, and we also calculated a simple energy correction based on these changes (Eisenberg et al., 1989). Changes in exposure of the various types of surface area due to complex formation were multiplied by atomic solvation parameters as described by Eisenberg *et al.* (1986, 1989), summed, and then added to our original interaction energies. The sole S atom in ubiquitin was treated as a polar N/O-type atom. This simple correction applies an energetic penalty for burial of polar or charged surfaces and an energetic reward for the burial of hydrophobic surfaces.

2.3 Results and Discussion

2.3.1 Relevant biochemical information

Ubiquitin is a highly conserved protein - the sequences of all animal ubiquitins are identical while yeast and plant ubiquitin each have three conservative substitutions (giving a total of four variant sites - residues 19, 24, 28, 57; Vierstra et al., 1986; Ozkaynak et al., 1984) . Yeast ubiquitin is fully active in assays of ubiquitin activation as well as ubiquitin-dependent proteolysis in animal-derived *in vitro* systems (Wilkinson et al., 1986). Oat ubiquitin is active in ubiquitin activation but stimulation of protein degradation has not been reported. It is expected to be fully active in this assay as well (Wilkinson, 1988). Wilkinson (1988) originally noted that the four variant residues of oat and yeast ubiquitin are clustered on one face of the protein and that this face, directly opposite to that of the carboxy-terminus, is probably not involved in intermolecular interactions in the ubiquitin-dependent proteolysis pathway. Subsequently, the crystal structure of diubiquitin revealed that all of these residues were distant from the dimer interface (Cook et al., 1992a). In the recently reported tetraubiquitin structure one of these variant residues (Glu 24) accepts two inter-monomer hydrogen bonds and another (Ala 28) is near an inter-monomer interface (Cook et al., 1994).

Wilkinson and co-workers have studied the effects of various chemical modifications of ubiquitin on ubiquitin activity in assays relevant to the ubiquitin-dependent proteolysis pathway (Wilkinson, 1988). Similarly, Ecker and co-workers

(Ecker et al., 1987) studied the effects of various mutations on the activity of ubiquitin in *in vitro* protein degradation. Although these studies do not provide direct evidence of the involvement of any specific residues or regions of the protein in a particular intermolecular interaction, they do hint at the relative importance or unimportance of certain residues in such interactions. We can use such suggestions as indicators of which sidechains might be involved in a protein-protein interaction at some point in the ubiquitin conjugation pathway. From their results we concluded that we should critically examine (see following section) residues Arg 42, 72, and 74, Tyr 59, and His 68. Unfortunately there was no such information available to us regarding mutants which could not be catalytically dimerized by ubiquitin conjugating enzyme.

From this variety of chemical and biochemical information (see above) we were able to construct a search space around the ubiquitin molecule which excluded the "variant face" of the docking target while at the same time allowing relatively unrestricted access of all possible orientations of the docking probe to a large part of the target surface. This accessible surface included all of the potentially critical residues described above. This search space excluded the possibility of obtaining dockings similar to the monomer-monomer configuration observed in the tetraubiquitin structure. However, the results of control experiments, as well as those of several analyses, suggest that the monomer-monomer configuration observed in tetraubiquitin is unlikely to be observed in a diubiquitin molecule (see below).

2.3.2 Relevant structural information

An obvious problem which can occur when using uncomplexed molecules to generate a complex during a rigid-body docking simulation is the clash of atoms which, in reality, could be avoided by very slight conformational adjustments (as observed, for example, in the diubiquitin structure - see Figure 2.4). The Lennard-Jones 6,12 potential used in our energy calculations ascribes prohibitive energy penalties to even slight atomic overlaps. In the context of the current work this means that, while the attraction due to any one sidechain in a large protein-protein interface can, in many cases, be omitted without significantly altering the dockings obtained, the repulsiveness of one unfavorably positioned sidechain can have a profound influence.

The flexibility of the C-terminal region of the ubiquitin molecule, described in Figure 2.1A, as well as the partial occupancy of these four residues, was originally noted by Vijay-Kumar and co-workers (Vijay-Kumar et al., 1987). For our rigid-body docking studies this is particularly challenging since this region of the molecule

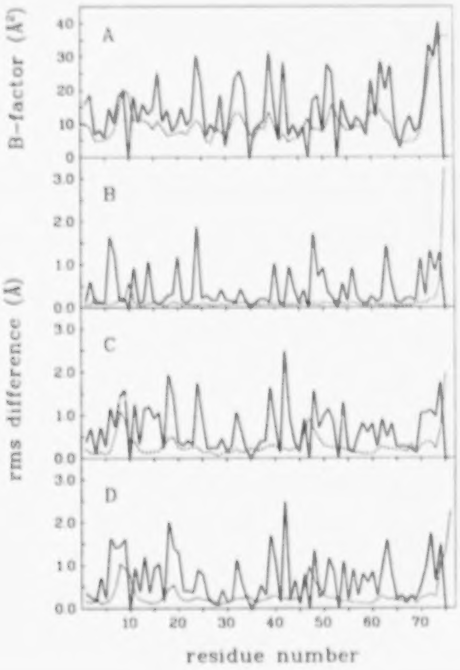


Figure 2.1: Indications of flexibility in ubiquitin and the two halves of diubiquitin. Solid lines, sidechain atoms; broken lines, backbone atoms. All superpositions were between the backbone atoms of residues 1-72. Backbone values for residue 76 (omitted from plots) are 5.5Å (B) and 4.8Å (C). (A) Average B-factors for the sidechain and backbone atoms of ubiquitin (1ubq). (B - D) RMS differences between the sidechains and backbones of the target and probe halves of diubiquitin (B), ubiquitin and the target (C), and ubiquitin and the probe (D).

is critical to the interaction of ubiquitin with molecules to which it becomes covalently attached. The combined backbone and sidechain flexibility in this region of the ubiquitin molecule allows for a prohibitive number of accessible conformational states. We did not attempt to model this flexibility directly in the docking simulations. Instead, we deleted Gly 76 prior to performing docking, in order to eliminate the possibility of a van der Waals clash between PROBE:76:C and TARGET:48:NZ during the docking simulation. This did not result in any major configurational changes to the complex upon rigid-body Monte Carlo minimization (probe 1 in Table 2.3).

Flexible sidechains of residues 1-72 include Glu 16, 24, and 64, Asn 25 and 60, Lys 33, Asp 39 and 52, Arg 42 and 72, and Gln 62 (Figure 2.1A). Several of these residues lie on the "variant face" of the target molecule which was excluded from our docking search (see above). These excluded residues include Glu 16 and 64, Asn 25 and 60, Lys 33, and Gln 62. Access to Glu 24 and 51 and Asp 52 was somewhat restricted. The only two relatively flexible sidechains (Figure 2.1A) in ubiquitin which were freely accessible to the probe molecule in our docking experiments were Asp 39 and Arg 42. In light of the difficulties we encountered when docking native (mono)ubiquitin (see below), it is interesting that of all of the relatively flexible sidechains in ubiquitin the difference between sidechain and backbone flexibility is greatest for Arg 42 (Figure 2.1A).

2.3.3 Docking - summary of experimental constraints

Prior to considering the information to be gained from the diubiquitin structure we summarize the salient biochemical and structural information and our application of it to the design of our docking simulations as follows. First, the variant residues of plant and yeast ubiquitin suggest that we can exclude this face of the ubiquitin molecule from our search. When we constructed a search cube that excluded this face of the protein we also excluded many of the flexible side chains in residues 1-72. This dramatically reduced the computational expense of our docking search and also eliminated many of the possible modifications which we might have considered (eg. multiple conformations, sidechain truncations). Second, structural information (cited above) indicated that flexibility in Asp 39, Arg 42 and 72, as well as residues 73-76 might create difficulties in our docking experiments. Chemical and biochemical information (cited above) had previously implicated several of these residues, as well as Tyr 59 and His 68, as being potentially critical in one or more protein-protein interactions involved in the ubiquitin-dependent protein degradation pathway. The search space we constructed, which excluded the "variant face" of the docking target, allowed relatively unhindered access of all possible orientations of the docking probe to the target surface comprising all of these critical residues. Third, in a wide variety of homodimeric proteins, the majority are found to exhibit two-fold symmetry (Miller, 1989). Our docking results were easily filtered to look for pseudo-twofold symmetric configurations. Fourth, we examined the nature (non-polar, polar, charged) of the interface surfaces in our dockings and compared these to the dimer interfaces previously characterized by other workers (Janin et al., 1988; Miller et al., 1987). Fifth, work by Chau and co-workers (Chau et al., 1989) as well as others (Gregori et al., 1990) has indicated that the most important ubiquitin-ubiquitin covalent linkage occurs between Lys 48 NZ of one monomer (the *target* in our experiments) and Gly 76 C of the second monomer (the *probe*). Again our docking results were easily filtered to look for dockings which would accommodate this constraint.

Of course, the crystallographically-observed structure of diubiquitin was available to us throughout the course of these docking experiments, and was in fact used to aid in construction of our reference complexes. However, in designing our docking experiments we attempted, as much as possible (see above), to use strategies that could have been deduced from previously available information, excluding the structure of diubiquitin itself. The information we used in our experimental design included the variant residues of ubiquitin, the *in vitro* effects of various chemical modifications to ubiquitin, as well as the dimensions of the ubiquitin monomer. Other information such as a covalent bond distance constraint and symmetry and surface considerations, as well as interaction energies and RMS differences between dockings and the appropriate reference structure, were used in the analysis of our docking results.

2.3.4 Docking - criteria of success

We consider an experiment to have been successful if the appropriate reference structure is generated during the docking search *and* that structure is ranked as the lowest energy docking by BOXSEARCH. Furthermore, we would like to see that the correct answer is a popular one - that is, if we group the dockings into clusters based on RMS differences the cluster containing the correct answer should be amongst the most heavily populated clusters. Since BOXSEARCH has been designed to generate all possible starting configurations with equal probability (Hart & Read, 1992), multiple visits to the more energetically favorable minima implies that our search has been reasonably exhaustive. We consider structures to be the same if the RMS difference between all atoms does not exceed 2 Å. Although our method allows for bias to be introduced prior to running the simulation, by modifying the molecules as well as by limiting the search space, once invoked the main docking algorithm itself is completely random and free of further bias [discussed in (Hart & Read, 1992)].

complex	$E^{a}_{complex}$ (kcal/mol)	E^{a}_{min} (kcal/mol)	RMS[b] (Å)	distance[c] bet. centres (Å)	rotation[c] angle (degrees)	translation[d] along screw axis (Å)	decrease in ASA[e] (Å^2)		
							nonpolar	polar	charged
Ub2	—	—	—	22.2	180.0	0.0	1033(.68)	240(.16)	243(.16)
Ub2/probe1	-75.7	-79.2	0.7	21.9	179.0	0.2	994(.67)	219(.15)	269(.18)
Ub2/probe2	-46.4	-68.6	0.5	22.1	179.3	0.1	878(.66)	216(.16)	242(.18)
mutant	+46.1	-39.0	2.0	23.7	176.6	1.8	641(.74)	158(.18)	72(.08)

Table 2.3: Reference structures for docking experiments. Data for the native diubiquitin complex (Ub2), two different diubiquitin complexes (probes 1 and 2), as well as our modified dimer (mutant). [a]Energies before and after minimization. [b]Movement caused by minimization. [c]Distances between the probe and target centres were measured after minimization, as were the rotations necessary to superpose the probe onto the target. [d]The minimized probes were superimposed onto their respective targets by rotation about an approximate twofold screw axis. This number represents the component of the translation parallel to the screw axis. [e]Changes in accessible surface area (ASA) were the differences between the complex and its two halves (fractions of the interface shown in parentheses).

2.3.5 Docking - reconstruction of diubiquitin from its two halves

The crystal structure of diubiquitin shows that the two ubiquitin monomers in this dimer are linked by an isopeptide bond between Gly A76 C and Lys B48 NZ (Cook et al., 1992a). Ubiquitin polymers consisting solely of Gly 76 - Lys 48 isopeptide bond-linked monomers have been shown to be fully competent mediators of ubiquitin-dependent proteolysis (Chau et al., 1989). While this is the most commonly observed linkage in various systems it is not the only one (Hochstrasser, 1992). The functions and relative importance of the various monomer-monomer linkages possible in ubiquitin polymers have yet to be determined. All of our docking simulations were aimed at generating dimers that might be covalently linked by an isopeptide bond between PROBE:76:C and TARGET:48:NZ.

We are unaware of any precise measurements of the affinity of the ubiquitin monomer for itself. If monoubiquitin self-associates in the absence of a conjugating enzyme we estimate a lower limit of 10 mM for the dissociation constant for non-covalent dimerization (calculation based on personal communication from M. Ellison). Obviously, the affinity of monoubiquitin for itself is low, at least when the monomers are not covalently linked. The two halves of diubiquitin are linked by a flexible chain that is potentially 20 Å in length when fully extended. This linkage allows for a variety of possible monomer-monomer interactions which combined encompass a relatively large configurational space. Conversely, Cook et al. (1992a) previously noted that it was possible to imagine a diubiquitin molecule in which the sole inter-monomer interaction was the covalent bond linking the two monomers.

Our first set of experiments dealt with the two halves of the crystallographically-observed diubiquitin structure. The only structural modification made in this case was the deletion of Gly 76 from both the target and the probe (probe 1). The interaction energy for this modified complex in the native configuration was calculated to be

-75.7 kcal/mol (Table 2.3). Rigid-body Monte Carlo energy minimization of this complex led to a 0.7 Å RMS shift and a slight decrease in the calculated interaction energy (Table 2.3). Initial experiments showed that the RMS differences between the dockings and our reference structures were significantly decreased when the reference configuration was minimized (results not shown).

When we ran an experiment with 5000 separate starts and an energy cutoff of -30.0 kcal/mol, 39 dockings fell below the energy cutoff (Table 2.4). Three of these dockings were correct with RMS differences from the reference structure of 0.6, 0.8, and 1.8 Å, and interaction energies of -74.5, -71.4, and -57.0 kcal/mol, respectively. The same experiment with 40000 starts produced 230 dockings below the energy

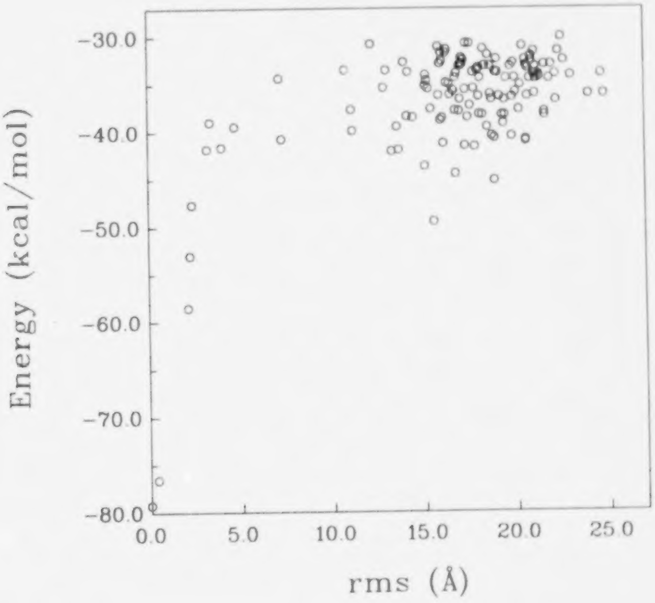


Figure 2.2: Reconstruction of diubiquitin from its two halves, judged by RMS deviation from a reference structure. 40000 separate docking starts with probe 1 gave rise to 230 dockings with interaction energies below -30.0 kcal/mol. These were separated into 131 clusters with the 3 lowest energy clusters containing a total of 28 dockings. The lowest energy member of each of the 131 clusters is shown in this figure. The point with an RMS value of 0.0 represents the minimized reference configuration.

cutoff (Table 2.4) and 21 of these were within 2 Å RMS of the reference structure. These 230 dockings were divided into 131 clusters (Figure 2.2), of which 86 had a single member. 28 of the 230 dockings fell into the 3 lowest energy clusters and all of the dockings that were within 2 Å RMS of the reference structure fell into the two lowest energy clusters (Table 2.5). Various statistics for this pair of experiments are shown in Tables 2.4 and 2.5. Figure 2.2 clearly shows that the dockings that are distant from the reference structure are energetically unfavorable relative to the correct dockings. Figure 2.3 compares the orientation of the lowest energy docking obtained in this experiment to that of the reference structure.

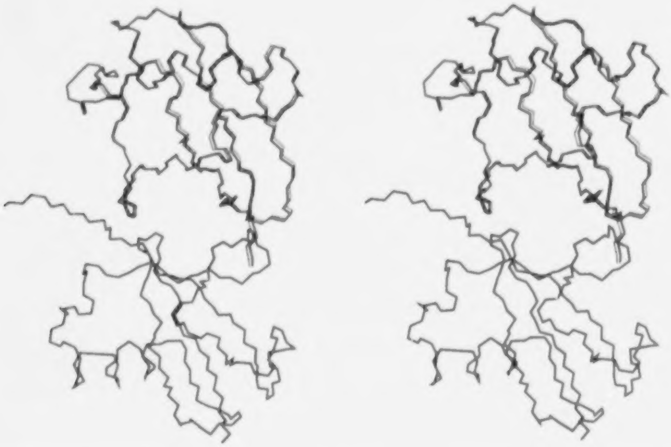


Figure 2.3: Superposition of the lowest energy docking (thin line) and the reference structure (thick line) for probe 1. The interaction energy of this docking was -76.6 kcal/mole and that of the reference was -79.2 kcal/mole. Only N, C, and CA atoms are shown.

According to all of the criteria outlined above this series of experiments was clearly successful. Not only is our docking protocol capable of generating dimer configurations within 2 Å RMS of the crystallographically-observed structure, it also ranks these dockings as the most favorable of all the dockings generated. Finally, cluster analysis of the dockings indicates that the correct docking is obtained relatively frequently. Indeed, in the large experiment described above, the low energy cluster has 3 times the number of members of the next most heavily-populated cluster. It is particularly exciting to note that in achieving this success we have not had to use all of the biochemical information available to us. Specifically, we have not filtered the results to remove dockings in which the C-terminus of the probe is distant from Lys 48 NZ of the target.

Our next set of experiments was a slightly more rigorous test of our docking protocol. Instead of docking together the two halves of diubiquitin (which may have undergone minor conformational changes to become more complementary), we used a copy of the target as the probe (probe 2). For a dimer which exhibits twofold pseudosymmetry, such as diubiquitin, we expect the interface regions of the two monomers to be quite similar to each other due to their similar environments. Since the dimer in this case is not perfectly symmetric, however, some differences between the two halves do exist. In general we would not expect the differences between the two halves of such a complex to be as great as the differences between the isolated monomer and either of the two halves of the complex. Figures 2.1B - 2.1D show that, with the exception of the C-terminus, the conformational differences between the two halves of diubiquitin are not as great as those observed for the ubiquitin monomer and the two halves of diubiquitin. This set of experiments gave us some insight into

# starts	# correct[a]	RMS[b] (Å)	energy[c] (kcal/mol)	cluster[d]
5000	3/39	0.6	-74.5	1/33
40000	21/230	0.4	-76.6	1/131

Table 2.4: Docking statistics for probe 1. The reference structure had an interaction energy of -79.2 kcal/mol. [a]Number correct / totai number of dockings which passed the energy cutoff. [b]Root-mean-square difference between the reference structure and the lowest energy docking. [c]Interaction energy of the lowest energy docking in that experiment. [d]Ranking of cluster containing best-fit answer / total number of clusters in this experiment.

the dependence of the success of our first set of experiments on a *particular* set of sidechain and C-terminus conformations.

cluster	energy[a] (kcal/mol)	RMS[b] (Å)	members[c]
1	-76.6	0.4	21
2	-58.5	2.1	5
3	-53.0	2.2	2
4	-49.5	15.5	5
5	-47.6	2.4	2
6	-45.2	18.8	5
7	-44.4	16.7	2
8	-43.6	15.1	4
9	-42.0	13.3	7
10	-41.9	13.6	5

Table 2.5: Top 10 clusters for large experiment with probe 1. Statistics listed are for the lowest energy docking in each cluster. The total number of dockings that passed the energy cutoff (-30.0 kcal/mol) in this experiment was 230. [a]Interaction energy of the lowest energy docking. [b]Root-mean-square deviation of the lowest energy docking from the reference structure. [c]Number of dockings in that cluster.

Despite the apparent decrease in favorability of this complex (Table 2.3) the results we obtained in this set of experiments were similar to those obtained in the first set of experiments (results not shown). Using this "modified probe" we were again able to generate and correctly rank the dockings which resembled the crystallographically-observed diubiquitin structure. The lowest energy cluster in the larger experiment was the most heavily populated cluster, and was represented by a docking which was within 3.1 kcal/mol and 0.5 Å RMS of the reference structure. The next best cluster was 14 kcal/mol less favorable, and was 17.4 Å RMS away from the reference configuration. Since the backbone conformations of the two halves of diubiquitin are virtually identical, with the exception of the C-terminus, this set of experiments showed that large modifications of the flexible parts of the probe (see

Figure 2.1B) did not prevent our docking protocol from generating and correctly scoring the experimentally observed diubiquitin structure. Also, once again we did not have to apply additional biochemical information during our analysis to achieve this success.

2.3.6 Docking - construction of diubiquitin with two copies of (mono)ubiquitin

A more rigorous and realistic test of a docking protocol is to attempt to reproduce the experimentally determined structure of a complex using the structures of the uncomplexed (*i.e.* native) components of the complex. This has been achieved in a number of cases (see, for example, Goodsell & Olson, 1990; Shoichet & Kuntz, 1991; Bacon & Moult, 1992; Hart & Read, 1992). Since most computational docking protocols, including our own, allow for little or no conformational flexibility in the interacting molecules, it is not surprising that docking results obtained with uncomplexed molecules are generally not as good as those obtained with the complex components (see, for example, Shoichet & Kuntz, 1991; Bacon & Moult, 1992; Hart & Read, 1992). In the absence of structural information of some sort the consideration of major (backbone) conformational changes which may be necessary for, or induced by, complex formation is problematic. This is especially true for the prediction of protein-protein complexes which may involve large interfaces and dozens of flexible sidechains. In the current study we chose to deal with the flexibility of certain critical residues in two ways - by systematically searching the accessible conformational states and, similar to Shoichet and Kuntz (1991), by truncating relatively disordered residues.

A summary of potentially relevant biochemical and structural information (see above) had indicated that the flexibility (or positioning) of Asp 39, Arg 42 and 72, as well as that of the C-terminal residues 73-76, might be crucial to this simulation. Our experiments with the different probes suggested that correct docking was not dependent on a *particular* C-terminus conformation (see above). In contrast to the comparison of the two halves of diubiquitin (Figure 2.1B), Figures 2.1C and 2.1D show that the conformations of Asp 39 and Arg 42 differ greatly between ubiquitin and the two halves of diubiquitin. Figure 2.1A shows that the sidechains of these two residues are amongst the most flexible in ubiquitin. Arginine residues are often among the most variable and uncertain in conformation. In principle, then, the potential importance of the conformations of these residues to the success of our docking simulations could be identified from any one of several lines of evidence.

Visual inspection of the dimer revealed that Arg 42 of both the target and the probe is located in the middle of the dimer interface. Figure 2.4 shows that while in the diubiquitin structure Arg 42 of the two halves easily accomodates dimer formation, the conformation of this sidechain in the monoubiquitin structure prohibits the formation of a diubiquitin-like dimer from two copies of monoubiquitin. A docking simulation with 2 copies of native ubiquitin, which was in all other respects identical to our

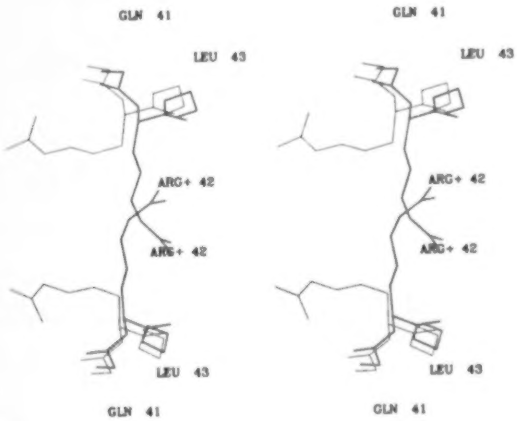


Figure 2.4: Superposition of two copies of monoubiquitin (thick lines) and the two halves of diubiquitin (thin lines). The different conformation of Arg 42, which is central to the diubiquitin interface, in monoubiquitin prevented dimer formation in our initial docking simulations with two copies of monubiquitin. The backbones of residues 41-43 and the Arg 42 sidechains are shown.

previous experiments, confirmed this (results not shown).

Modelling of sidechain flexibility during a docking simulation increases the difficulty of an already challenging problem and we did not wish to address this related issue in the current work. Instead, we took the very simplified approach of approximating the flexibility of the Arg 42 sidechains by truncating them down to Ala residues. Perhaps surprisingly, this worked.

Two copies of this modified ubiquitin molecule (Arg 42 → Ala 42; Gly 76 deleted) were superimposed onto the diubiquitin structure to generate a reference dimer configuration. The interaction energy of this unminimized configuration was +46.1 kcal/mol (Table 2.3). Rigid-body Monte Carlo minimization gave a dimer configuration with a reduced interaction energy (Table 2.3). This value was still somewhat higher than those observed in our earlier experiments (Table 2.3). Although the RMS difference between this minimized probe and the unminimized probe was relatively large when compared to the values obtained in our earlier experiments (2.0 Å *versus* 0.7 or 0.5 Å; Table 2.3, several other statistics indicated that it represented a dimer configuration which was similar to the reference complexes we had used in our earlier experiments (Tables 2.3 and 2.6). Since this experiment involved the uncomplexed monomer the less favorable values we observed were not surprising. With the exception of the Arg 42 → Ala 42 modification, the protocol of this docking simulation was identical to that employed in our previous experiments.

In the experiment with the modified ubiquitin monomer, 40000 starts yielded 184 dockings with interaction energies below -30.0 kcal/mol. These divided into 110 clusters, 73 of which contained a single member (Figure 2.5). Figure 5 shows that 2 of the 3 lowest energy clusters are within 4.6 Å RMS of the reference structure while the other low energy cluster is 23.1 Å RMS away. The lowest energy dockings of clusters 1 and 2 are shown in Figure 2.6. We see that these two dockings utilize radically different interfaces; these differences are further detailed by the

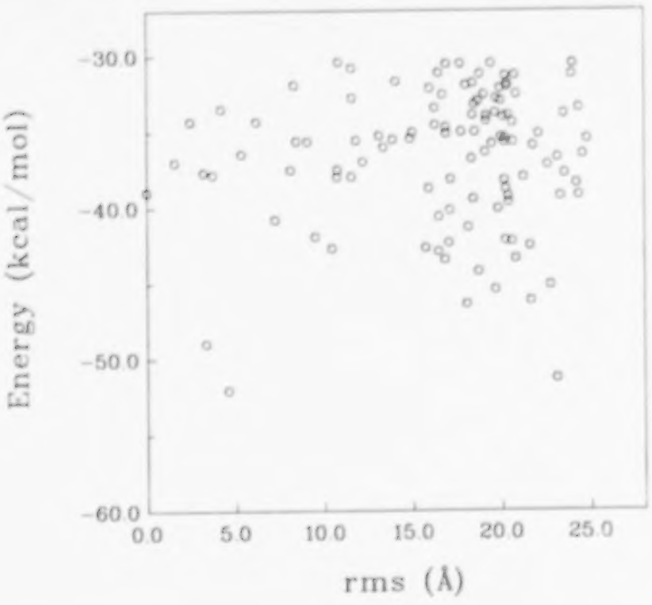


Figure 2.5: Reconstruction of diubiquitin from "mutant" ubiquitin, judged by RMS deviation from a reference structure. 40000 separate docking starts with two copies of our modified ubiquitin molecule gave rise to 184 dockings with interaction energies below -30.0 kcal/mol. These were separated into 110 clusters with the 3 lowest energy clusters containing a total of 16 dockings. The lowest energy member of each of the 110 clusters is shown in this figure. The point with an RMS value of 0.0 represents the minimized reference configuration.

measurements presented in Table 2.6 (discussed below). 14 of the 184 dockings are contained in the 3 lowest energy clusters, and cluster 1 is by far the most heavily populated cluster in this experiment (Table 2.6). Figure 2.5 also shows that many dockings, both near to and distant from the reference structure, have lower interaction energies than the reference structure. In contrast to our experiments involving the two halves of diubiquitin, as well as those involving two copies of one half of diubiquitin, this docking experiment was not unambiguously successful when judged solely on the basis of docking energies and RMS differences. None of the top 10 clusters are within 2 Å RMS of our reference structure. Furthermore, the top 10 clusters are all represented by dockings of lower energy than the reference structure. However,

cluster	energy^a	adjusted^b	RMS^c	members^d	pseudo-bond^e	rotation onto^f		translation^g	rotation onto^h		decrease in ASA^i (Å^2)		
	(kcal/mol)	energy	(Å)		distance (Å)	target		along screw	probe		nonpolar	polar	charged
		(kcal/mol)				angle (deg)	distance (Å)	axis (Å)	angle (deg)	distance (Å)			
Ub2/probe1	-79.2	-85.0	—	—	4.3	180.0	22.2	0.2	—	—	1033(.68)	240(.16)	243(.16)
mutant	-39.0	-46.4	0.0	—	5.7	176.6	23.6	1.8	0.0	0.0	641(.74)	158(.18)	72(.08)
1	-52.1	-59.3	4.6	11	4.1	179.3	23.4	0.0	14.6	3.7	738(.69)	224(.21)	110(.10)
2	-51.3	-43.6	23.1	2	24.1	162.6	22.3	3.6	140.5	15.4	693(.51)	165(.12)	499(.37)
3	-49.0	-56.9	3.3	3	5.2	178.9	23.5	0.1	11.2	2.6	748(.72)	195(.19)	101(.10)
4	-46.4	-42.2	18.0	3	13.9	174.0	23.0	4.9	85.8	11.8	494(.49)	222(.22)	296(.29)
5	-46.2	-38.9	21.6	4	22.8	133.1	23.5	10.7	153.1	12.7	394(.46)	93(.11)	362(.43)
6	-45.4	-40.0	19.6	1	15.7	174.4	23.0	5.3	95.1	13.0	497(.48)	213(.20)	331(.32)
7	-45.1	-39.8	22.7	1	23.3	162.5	23.5	3.2	141.1	15.0	620(.54)	114(.10)	412(.36)
8	-44.2	-39.7	18.7	1	19.6	160.6	22.9	5.7	98.3	10.6	526(.51)	183(.18)	328(.32)
9	-43.4	-42.5	16.8	3	23.7	153.0	23.8	21.7	150.5	1.8	489(.54)	195(.22)	212(.24)
10	-43.4	-37.4	20.8	3	23.1	126.6	23.4	10.7	157.7	11.0	407(.48)	100(.12)	331(.40)

Table 2.6: Top 10 clusters for experiment with modified ubiquitin. Statistics listed are for the lowest energy docking in each cluster. The total number of dockings that passed the energy cutoff (-30.0 kcal/mol) in this experiment was 184. ^{a}See Table 2.5. ^{b} A simple correction for solvation effects was made (see Methods). ^{c,d}See Table 2.5. ^{e}PROBE:75:C to TARGET:48:NZ distance. ^{f}Rotation necessary to superimpose the docking onto the target and the distance between centres of the two molecules. ^{g}See Table 2.3. ^{h}Rotation necessary to superimpose the docking onto the reference probe and the distance between centres of the two molecules. ^{i}See Table 2.3.

by exploiting additional biochemical and structural information, as discussed below, the ambiguity can be removed. We describe several analyses that together clearly indicate that clusters 1 and 3 are essentially correct dockings. It is thus possible to predict clearly a diubiquitin-like dimer from two copies of the modified monomer, using information derived independently of the diubiquitin structure.

We applied Eisenberg and McLachlan's (1989) correction for solvation effects (not considered in our current energy calculation) to the calculated interaction energies and this improved the relative ranking of the top 10 clusters such that only clusters 1 and 3 were of lower energy than the reference structure (Table 2.6). This is a reflection of the nature and extent of the buried surface in each of the dockings (Table 2.6). Janin (Janin et al., 1988) and Miller (Miller et al., 1987) have compiled a detailed summary of the nature of the accessible surfaces of both monomers and polymers. Both ubiquitin and diubiquitin fit the description given by these authors regarding the nature of accessible surface area as well as that of the dimer interface. Of the 18 dimer interfaces studied by these authors (Janin et al., 1988) none were more than approximately 22% charged or 30% polar. The information regarding charged interface surface area is compelling. While both diubiquitin and our reference dimer fall within the boundaries outlined by these authors, all of the top 10 clusters, with the exception of clusters 1 and 3, have relatively high proportions of buried charged surfaces in the dimer interface (Table 2.6). The application of this type of information as a data filter is a considerable aid in the analysis of (the huge amount of) data obtained in these docking simulations. The potential function used in the current docking simulations does not consider solvent effects at all and these results clearly indicate the need for, and potential utility of, such a term in our energy calculation. On the other hand, the dramatic improvement we obtained by applying this solvation correction to our docking results is in contrast to the reports of other groups. Shoichet & Kuntz (1991) reported docking results of both bound and unbound molecules for

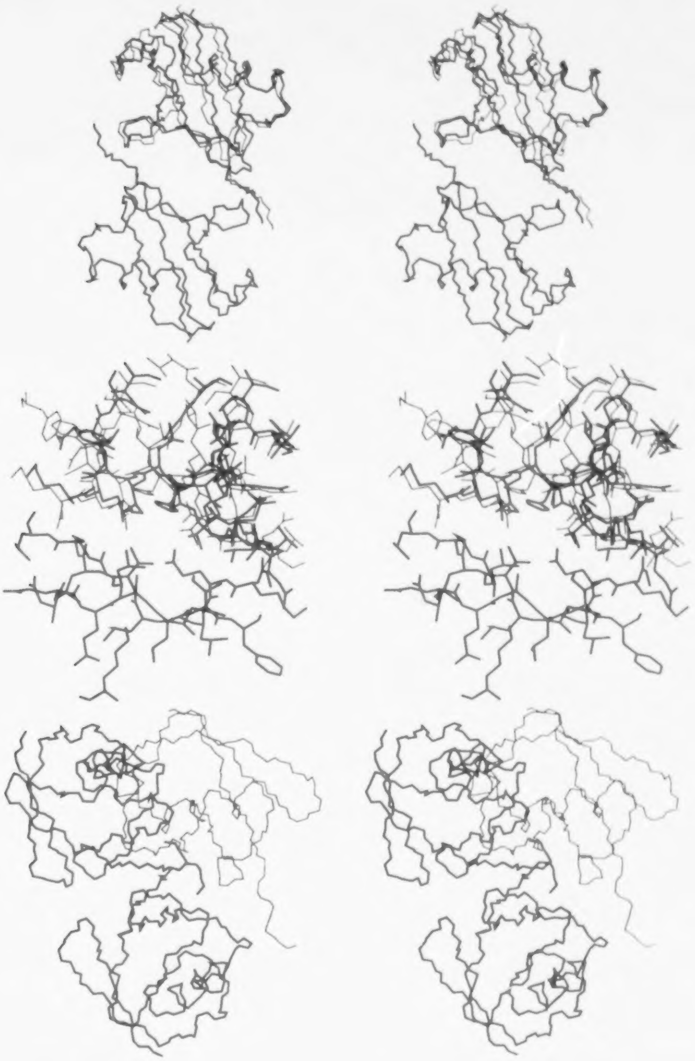


Figure 2.6: Docking with monoubiquitin. Each stereo pair shows a target (lower; thick line) as well as a superposed reference (upper; thick lines) and docking (upper; thin lines). *Top:* N, C, CA atoms of the lowest energy docking in cluster 1 (interaction energy = -52.1 kcal/mole; see Table 2.6); *middle:* Details of the interface of the docking shown above. In this figure all heavy atoms of all residues containing an atom within 12Å of CB of the modified (Arg 42 → Ala 42) residue (in the reference probe) are shown; *bottom:* N,C, CA atoms of the lowest energy docking in cluster 2 (interaction energy = -51.3 kcal/mole; see Table 2.6).

3 different systems. Applying a similar correction to the one we used (compare the

ASP values of Eisenberg & McLachlan, 1986, with those of Eisenberg et al., 1989) they achieved no significant improvement in the relative energy rankings of any of their reported dockings. It is possible that these authors might have achieved greater success with this approach if they had used the parameter set employed in the present work. Alternatively, these discrepancies may indicate a lack of general applicability of this method of correcting for solvation effects. Detailed studies aimed at addressing this question are currently underway in this laboratory (see Chapter 3). Janin and co-workers (Cherfils et al., 1991; Cherfils et al., 1994) found no correlation between docking "correctness" and total buried surface area. These latter authors did not attempt to distinguish between various surface types in their calculations.

A mechanistic or functional analysis of the dockings also proved to be quite useful. Several other groups have derived distance constraints from a variety of non-energetic information, and the application of such distance filters has been shown to simplify the analysis of a variety of docking results (Cherfils et al., 1991; Shoichet & Kuntz, 1991; Stoddard & Koshland, 1992; Cherfils et al., 1994). Since we were interested in a diubiquitin structure linked by an isopeptide bond between the C-terminus of the probe and Lys 48 of the target we measured a representative distance for the top 10 clusters as well for two of our reference structures (Table 2.6). Since in all of our experiments we deleted residue 76, we measured the distance between PROBE:75:C and TARGET:48:NZ, in the reference structure(s) as well as the dockings, to determine the feasibility of isopeptide bond formation between the probe and the target (Table 2.6). Neglecting the possibility of large conformational shifts, isopeptide bond formation between the two molecules is only possible for clusters 1 and 3. Since the C-terminus of the probe as well as Lys 48 of the target are both flexible we further explored the possibility of bond formation by running a systematic conformational search on these flexible regions of the molecules. A docking from an earlier experiment which was similar (0.7 Å RMS) to the low energy docking of cluster 2 had its C-terminus trimmed from LeuArgGlyGly to AlaAlaGlyGly to simulate sidechain flexibility. We then systematically searched the accessible conformational space of these 4 phi-psi pairs, as well as the 4 sidechain torsion angles of Lys 48 of the target, in 30° steps. With scaled down van der Waals radii the closest conformer had a PROBE:75:C to TARGET:48:NZ distance of 9.0 Å. Using the native sequence (sidechains of Leu 73 and Arg 74 fixed) we obtained no conformers in which this distance was less than 12 Å. In contrast, when we searched the 4 sidechain angles of Lys 48 and *only the last* phi angle in the C-terminus of the probe in our reference configuration we obtained 408 conformers in which the distance of interest was between 2 and 4 Å.

We also examined these docking results for configurations representative of dimers which could be covalently linked *via* one of the other Lys residues of the target. Very few of the low energy dockings had pseudo-bond distances which would allow for covalent bond formation between the C-terminus of the probe and any of these other Lys residues of the target without major conformational adjustments. The lowest such distance for cluster 2 (Table 2.6) was 13.1 Å for Lys 27 of the target. Cluster 8 (Table 2.6) had a pseudo-bond distance of 7.5 Å with Lys 6 of the target, and cluster 19 had an equivalent distance of 9.1 Å. With these two exceptions the most energetically-favorable dockings which meet this covalent constraint are those represented by clusters 1 and 3. The most likely covalent linkage for these dockings is quite clearly the same as that observed in the diubiquitin structure. With information of this sort available this type of data filtering would be of obvious value in a *real* prediction situation where a reference structure is not available.

That a homodimer will generally exhibit two-fold symmetry, or at least pseudo-symmetry, was first predicted by Monod (Monod et al., 1965) and is supported by the empirical work of Miller (Miller, 1989) and others (see refs in Miller, 1989). We measured the rotation and translation necessary to superimpose our dockings onto the target molecule. Diubiquitin exhibits twofold pseudosymmetry and our reference dimer (modified) is similar (Table 2.6). Clusters 1 and 3 have much closer twofold pseudosymmetry than most of the other clusters (Table 2.6). We used a method similar to that described by Shoichet and Kuntz (Shoichet & Kuntz, 1991) to measure the difference between the dockings and the reference probe in terms of rotation angle and translation distance. Table 2.6 shows that this measurement approximately parallels the ranking according to the RMS differences between the dockings and the reference probe.

One final point worthy of mention is our criterion of which structures are the same. All of our analyses have been based on the arbitrary assumption that structures within 2 Å RMS of each other are the same whereas more distant structures are different. For complexes involving two large molecules, particularly if the complex is of relatively low affinity, this may be an unrealistically limiting criterion. Indeed, our preliminary findings in this area with both gradient and Monte Carlo minimization suggest that more distant (than 2 Å RMS) configurations often converge to the same minimum (results not shown). Also, the relationship between the diubiquitin and tetraubiquitin (see below) structures supports the idea that diubiquitin has considerable configurational adaptability in solution.

2.3.7 Docking - tetraubiquitin

While the present manuscript was in preparation the structure of tetraubiquitin was reported (Cook et al., 1994). In contrast to the previously reported diubiquitin structure which is the focus of the current work, the ubiquitin-ubiquitin interactions in tetraubiquitin allow for indefinite extension of the ubiquitin polymer along a twofold screw axis. As a docking problem the prediction of the ubiquitin-ubiquitin configuration observed in the tetraubiquitin structure is much more difficult because a given ubiquitin monomer interacts with more than one other ubiquitin monomer.

We were, of course, interested in re-examining our results in light of this new information. A comparison of diubiquitin (only Gly 76 deleted; probe 1 in Table 2.3) and the appropriate dimer from the tetraubiquitin structure was most telling. With our potential function we calculated interaction energies of -75.7 kcal/mol for diubiquitin and -9.7 for the tetraubiquitin dimer. The corresponding values corrected for solvation effects (see above) were -81.4 and -4.9 kcal/mol, respectively. Upon minimization the diubiquitin structure shifted 0.7 Å RMS and the energy decreased slightly to -79.2 kcal/mol. Minimization of the tetraubiquitin dimer produced a more dramatic shift of 3.4 Å RMS and a new interaction energy of -33.4 kcal/mol. Solvation correction of these latter two interaction energies gave values of -85.0 and -28.1 kcal/mol, respectively. The interface area of the tetraubiquitin dimer is relatively small (595 Å^2) and the proportion of charged area is very high (45%) when compared to other dimers (Janin et al., 1988; Miller et al., 1987) as well as our reference diubiquitin configurations and dockings (Tables 2.3 and 2.6).

Superimposing the target of the appropriate dimer from the tetraubiquitin structure onto the target in our docking simulation revealed that the probe from the new configuration extended, unfortunately, approximately 4 Å beyond the search space of our simulations. We would not, therefore, have found this configuration in our earlier docking experiments. Using a slightly larger search space (56 Å cube) placed so as to easily accommodate the new configuration (shifted 8 Å along one axis), as well as our earlier results, we constructed a reference "tetraubiquitin dimer" by superimposing our modified target and probe (see preceding section; Arg 42 → Ala and Gly 76 deleted) onto the appropriate halves of a Gly 76 - Lys 48 isopeptide-linked pair from the tetraubiquitin structure. Rigid-body Monte Carlo minimization of this configuration resulted in a relatively large shift of 5.6 Å RMS. The original configuration had an interaction energy of +143.0 kcal/mol; minimization reduced this to -35.0 kcal/mol, so this reference structure did pass the (arbitrarily chosen) energy cutoff employed in our earlier simulations. The difference between the effects of minimization on this dimer and the native tetraubiquitin dimer (this dimer shifted 2.2 Å RMS more than the native dimer; see preceding paragraph) could not be ascribed to one or a few particular sidechain conformations or steric clashes. It is likely a reflection of the unsuitability of this interface for a simple monomer-monomer interaction. When we ran a docking simulation with this dimer in the new search space (see above) with 40000 starts, 162 dockings passed the energy cutoff of -30.0 kcal/mol (results not shown). One relatively high energy docking was 4.2 Å RMS away from the minimized reference probe; no other dockings were within 11 Å RMS of this reference structure. On the other hand, several diubiquitin-like configurations *were* obtained, and some of these were amongst the lowest energy configurations observed. The most energetically favorable of these dockings was 3.5 Å RMS away from the minimized diubiquitin reference and had an interaction energy of -49.0 kcal/mol. The PROBE:75:C to TARGET:48:NZ distance of this docking was 5.2 Å, similar to that of the reference configuration (see above and Table 2.6). 1 of the 8 dockings of lower energy (-50.2 kcal/mol) which were obtained in this simulation also had a favorable PROBE:75:C to TARGET:48:NZ distance (4.8 Å). This docking was 5.0 Å RMS away from the minimized diubiquitin reference configuration. The average PROBE:75:C to TARGET:48:NZ distance of the other 7 lower energy dockings was 18.8 Å; the smallest was 14.8 Å. Without major conformational changes covalent bond formation between the two molecules seems possible for only two of these low energy dockings.

This docking result suggested that in the absence of further interactions, such as those observed in tetraubiquitin, and also, presumably, in higher order ubiquitin polymers, the dimer configuration observed in tetraubiquitin is not particularly favorable, at least according to our potential function. To further explore this issue we ran a docking simulation in which we attempted to reassemble the two unmodified halves of a dimer taken from the tetraubiquitin structure (this experiment was analogous to our first two diubiquitin experiments with probe 1). In this case very few dockings passed the -30.0 kcal/mol energy cutoff and none were within 15 Å RMS of the native or minimized reference probe. This result offers further support for the contention that for a Gly 76 - Lys 48 isopeptide-linked ubiquitin *dimer* the monomer-monomer interaction observed in the tetraubiquitin structure is not particularly stable.

Taken together our results indicate that a Gly 76 - Lys 48 isopeptide-linked ubiquitin *dimer* can find more favorable interactions than those present between two adjacent monomers in the tetraubiquitin structure. The crystallographically-observed diubiquitin structure is an example of a more favorable ubiquitin-ubiquitin interaction, and may represent the most favorable interaction for such a covalently linked pair. In discussing the tetraubiquitin structure the authors (Cook et al., 1994) conclude that the diubiquitin structure probably represents the predominant form of ubiquitin in solution and that the tetraubiquitin configuration is adopted when a third monomer is ligated to diubiquitin. Our docking simulations, as well as our energy and surface area calculations, are consistent with this conclusion.

2.3.8 Docking - summary

Reconstruction of a crystallographic complex is the standard test of a docking protocol. Several different methodologies, including the one employed here, have passed this test in studies with a variety of biochemical systems (refs cited above). A more rigorous and realistic test of a docking protocol is to attempt to reproduce the experimentally determined structure of a complex using the structures of the uncomplexed (*i.e.* native) components of the complex. This has also been achieved in a number of cases (refs cited above).

The present work differs significantly from previous examples, however, in a number of ways. First, we are studying a system that, to our knowledge, has not been investigated by such methods. Although we are predicting answers which have already been determined experimentally, our primary interest is in the prediction of the currently unknown structures of complexes involving ubiquitin and enzymes of the ubiquitin conjugation pathway. The quality of the results reported here indicates to us that we may be able to make such predictions reliably. Second, ubiquitin interacts with a variety of non-homologous enzymes (Rechsteiner, 1991; Hershko, 1991; Finley & Chau, 1991; Hochstrasser, 1992; Hershko & Ciechanover, 1992; Jentsch, 1992). Most docking studies have focussed on target-probe interactions that are very specific and of relatively high affinity (see, for example, Goodsell & Olson, 1990; Shoichet & Kuntz, 1991; Bacon & Moult, 1992; Hart & Read, 1992). We are encouraged by our ability to predict the stucture of a complex involving this "indiscriminate" protein. Third, we have applied a variety of non-energetic biochemical information to the analysis of our docking results in a systematic, quantitative, and productive manner. While most of the methods we applied have been reported previously, the variety of information we found to be applicable to this problem, as well as the extent to which these filters clarified the analysis of the results of our final docking simulation, is particularly encouraging. Fourth, most docking studies have investigated non-covalent complexes. It is unknown what part, if any, non-covalent intermolecular interactions play in the formation or stabilization of covalent ubiquitin complexes. As Shoichet & Kuntz (1991) have previously observed, the existence of a covalent bond between the two components of a complex can potentially complicate, as well as simplify, a docking study. Our docking simulations were performed without consideration of covalent bond formation between the two ubiquitin subunits (except, in one case, for filtering the docking results). The results reported here suggest that non-covalent intermolecular interactions are important for the formation and/or stabilization of the crystallographically observed diubiquitin complex.

It might be argued that ubiquitin self-associates too weakly to provide a good system for docking studies (we estimate a lower limit for K_d of 10 mM - see above). Nonetheless, it is not surprising that the non-covalent affinity is low in a complex with a covalent bond, since evolution will only proceed to the point that there is a moderate energy stabilizing the desired configuration(s). On the other hand, a requirement for specificity means that the energy *difference* between the desired configuration(s) and all other possibilities must be large compared to kT. The success of a docking experiment depends on the discrimination of energy differences of this size, and not on absolute binding energies.

Another aspect of the present work is the development of general docking strategies. In this respect the experiments reported here serve several purposes. First, while sidechain flexibility is crucial, at least in some cases, to successful docking, we report several more examples of the effectiveness of a relatively crude approximation of this flexibility, residue truncation. Second, the consideration of only non-covalent interactions during docking can lead to correct predictions with a covalent complex. Third, and perhaps most important, the consideration of diverse chemical and biochemical information can dramatically clarify the results obtained from docking simulations. Fourth, we have seen that certain modifications of our docking procedure, such as the incorporation of a surface burial term and a different minimization scheme, could increase the power of that procedure.

We have discussed the limitations imposed by rigid-body docking and the difficulty of allowing for major conformational changes during docking simulations (see above). Sidechain flexibility, on the other hand, can be modelled during docking simulations. In the current study we chose to deal with the flexibility of certain critical residues in two ways - by systematically searching the accessible conformational states of a docked complex and by truncating relatively disordered residues prior to docking.

Conformational searching proved to be a powerful way of incorporating biochemical information into the analysis of our docking results. Truncation, obviously, is a radical approximation of flexibility and by no means ideal, especially when the residue of interest is part of the intermolecular interface involved in the docking study. When the intermolecular interaction is between two proteins and involves a large interface the truncation of one or two sidechains may remove a prohibitive steric clash without otherwise affecting the association. A better approximation would be to include a limited rotamer library of flexible sidechains which could be sampled during the docking simulation. The application of the dead-end elimination theorem to the prediction of sidechain conformation has recently been described (Desmet et al., 1992), and its incorporation into a docking protocol has been reported (Leach, 1994). This method is also based on a library of allowed sidechain rotamers. The judicious implementation of some type of discrete-sampling approach, based on a user-defined rotamer library, to address the problem of sidechain flexibility in docking seems to be computationally feasible at this time and we plan to incorporate such an improvement in our method.

The ideal docking experiment would search all possible configurations of the complex of interest and pick the correct one to be the one of lowest energy. Furthermore, this conclusion would be arrived at without considering any additional information (*eg.* binding or mutation studies, chemical modifications). Current methods do not allow for this ideal experiment due to a variety of limitations.

Drug design is one of the common applications of docking simulations. Consideration of a typical drug design scenario, however, leads one to the conclusion that such a powerful method is not strictly necessary (although it is, of course, desirable). Simply speaking, in this scenario the investigators will have a target structure derived from either experiment or calculation, several structurally-related ligands which exhibit a wide range of affinities for the target, and some information regarding the nature of the site of interaction (from, for example, competition or mutation studies). This type of information can be incorporated into a docking study to greatly reduce the configurational space which must be searched. This in turn will allow for a more exhaustive search of the smaller space and will increase the chances of determining the correct binding mode(s). Alternatively, such information can be applied as a filter to reject some of the data obtained in an unrestricted docking simulation. Variations of these approaches have been reported by several other groups (refs cited above). We have successfully applied both of these approaches in the present studies of the diubiquitin system.

2.4 Conclusion

Upon consideration of the limitations and approximations involved in current docking simulations, indeed in simulations of biomolecules in general, surprise at the quality of the results obtained in many of these simulations is, perhaps, justified. While the state of the art of simulations continues to evolve towards a truer representation of reality, the disparity that exists between current ideals and implementations will undoubtedly persist for several years. Irrespective of this, many workers continue to achieve success in the field.

In the current work we have applied our particular implementation of a method to solve the docking problem to a new and challenging biochemical system. Consideration of biochemical and structural information derived from a variety of sources and the application of such to both experimental design and the analysis of our results has allowed us to generate and correctly score diubiquitin-like dimer configurations using the two halves of diubiquitin, two copies of one of the halves of diubiquitin, as well as two copies of a modified form of the uncomplexed ubiquitin monomer. Docking results, as well as the results of surface area and energy calculations, are consistent with the observation of distinct configurations for a ubiquitin dimer and tetramer. The monomer-monomer interaction observed in tetraubiquitin is relatively unfavorable for a simple covalently linked ubiquitin dimer. Our ability to predict the crystallographically-observed dimer configuration supports the idea that this structure represents the biologically relevant dimer configuration.

Finally, we briefly discuss the evolution of our docking method(s). One limitation of current concern is the inefficiency of Monte Carlo minimization in getting to the bottom of local minima. In the near future our method will be modified to include a two stage minimization scheme involving an initial stage of Monte Carlo search followed by a final stage of gradient minimization. Our preliminary investigations in this area indicate that the clusters will become much tighter, reducing the complexity of the results obtained from docking simulations. Also, while we consider our docking simulations to have been successful in the three systems described here, as well as those described elsewhere (Hart & Read, 1992; Hart & Read, 1994), the data presented clearly show that incorporation of a surface burial term in our potential function would help to clarify the docking results obtained, at least in this case. This observation is not surprising and we plan to incorporate a term to account for solvation effects in the next version of BOXSEARCH. Finally, we will be introducing methods for more accurately modelling flexibility as well as for incorporating relevant biochemical and structural information into our simulations.

2.5 References

Bass, M. B., Hopkins, D. F., Jaquysh, W. A. N., & Ornstein, R. L. (1992). A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. *Proteins: Structure, Function, and Genetics*, 12:266–277.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542.

Chau, V., Tobias, J. W., Bachmair, A., Marriott, D., Ecker, D. J., Gonda, D. K., & Varshavsky, A. (1989). A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. *Science*, 243:1576–1583.

Chen, Z. & Pickart, C. M. (1990). A 25-kilodalton ubiquitin carrier protein (E2) catalyzes multiubiquitin chain synthesis via lysine 48 of ubiquitin. *J. Biol. Chem.*, 265:21835–21842.

Cherfils, J., Bizebard, T., Knossow, M., & Janin, J. (1994). Rigid-body docking with mutant constraints of influenza hemagglutinin with antibody HC19. *Proteins: Structure, Function, and Genetics*, 18:8–18.

Cherfils, J., Duquerroy, S., & Janin, J. (1991). Protein-protein recognition analyzed by docking simulation. *Proteins: Structure, Function, and Genetics*, 11:271–280.

Cherfils, J. & Janin, J. (1993). Protein docking algorithms: simulating molecular recognition. *Curr. Op. Struc. Biol.*, 3:265–269.

Cook, W. J., Jeffrey, L. C., Carson, M., Chen, Z., & Pickart, C. M. (1992a). Structure of a diubiquitin conjugate and a model for interaction with ubiquitin conjugating enzyme (E2). *J. Biol. Chem.*, 267:16467–16471.

Cook, W. J., Jeffrey, L. C., Kasperek, E., & Pickart, C. M. (1994). Structure of tetraubiquitin shows how multiubiquitin chains can be formed. *J. Mol. Biol.*, 236:601–609.

Cook, W. J., Jeffrey, L. C., Sullivan, M. L., & Vierstra, R. D. (1992b). Three-dimensional structure of a ubiquitin-conjugating enzyme (E2). *J. Biol. Chem.*, 267:15116–15121.

Cook, W. J., Jeffrey, L. C., Xu, Y., & Chau, V. (1993). Tertiary structures of class I ubiquitin-conjugating enzymes are highly conserved: Crystal structure of yeast UBC4. *Biochemistry*, 32:13809–13817.

Desmet, J., De Maeyer, M., Hazes, B., & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542.

Deveraux, Q., Ustrell, V., Pickart, C., & Rechsteiner, M. (1994). A 26 S protease subunit that binds ubiquitin conjugates. *J. Biol. Chem.*, 269:7059–7061.

Ecker, D. J., Butt, T. R., Marsh, J., Sternberg, E. J., Margolis, N., Monia, B. P., Jonnalagadda, S., Khan, M. I., Weber, P. L., Mueller, L., & Crooke, S. T. (1987). Gene synthesis, expression, structures, and functional activities site-specific mutants of ubiquitin. *J. Biol. Chem.*, 262:14213–14221.

Eisenberg, D., Wesson, M., & Yamashita, M. (1989). Interpretation of protein folding and binding with atomic solvation parameters. *Chemica Scripta*, 29A:217–221.

Evans, S. V. (1993). Setor: hardware lighted three-dimensional solid model representations of macromolecules. *J. Mol. Graphics*, 11:134–138.

Finley, D. & Chau, V. (1991). Ubiquitination. *Annu. Rev. Cell Biol.*, 7:25–69.

Gregori, L., Marriott, D., West, C. M., & Chau, V. (1985). Specific recognition of calmodulin from *Dictyoselium discoideum* by the ATP,ubiquitin-dependent degradative pathway. *J. Biol. Chem.*, 260:5232–5235.

Gregori, L., Poosch, M. S., Cousins, G., & Chau, V. (1990). A uniform isopeptide-linked multiubiquitin chain is sufficient to target substrate for degradation in ubiquitin-mediated proteolysis. *J. Biol. Chem.*, 265:8354–8357.

Hart, T. N. & Read, R. J. (1992). A multiple-start Monte Carlo docking method. *Proteins: Structure, Function, and Genetics*, 13:206–222.

Hart, T. N. & Read, R. J. (1994). Multiple-start Monte Carlo docking of flexible ligands. In Merz, K.M., J. & LeGrand, S., editors, *The Protein Folding Problem and Tertiary Structure Prediction*, pages 71–108. Birkhäuser, Boston.

Hershko, A. (1991). The ubiquitin pathway for protein degradation. *TIBS*, 16:265–268.

Hershko, A. & Ciechanover, A. (1992). The ubiquitin system for protein degradation. *Annu. Rev. Biochem.*, 61:761–807.

Hershko, A. & Heller, H. (1985). Occurrence of a polyubiquitin structure in ubiquitin-protein conjugates. *Biochem. Biophys. Res. Commun.*, 128:1079–1086.

Hill, C. P., Johnston, N. L., & Cohen, R. E. (1993). Crystal structure of a ubiquitin-dependent degradation substrate: A three-disulfide form of lysozyme. *Proc. Nat. Acad. Sci.*, 90:4136–4140.

Hochstrasser, M. (1992). Ubiquitin and intracellular protein degradation. *Curr. Op. Cell Biol.*, 4:1024–1031.

Janin, J., Miller, S., & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.*, 204:155–164.

Jentsch, S. (1992). The ubiquitin-conjugation system. *Annu. Rev. Genet.*, 26:179–207.

Kuntz, I. D., Meng, E. C., & Shoichet, B. K. (1994). Structure-based molecular design. *Acc. Chem. Res*, 27:117–123.

Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.*, 235:345–356.

Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins. *Prot. Eng.*, 3:77–83.

Miller, S., Janin, J., Lesk, A. M., & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.*, 196:641–656.

Monod, J., Wyman, J., & Changeux, J. (1965). On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.*, 12:88–117.

Ozkaynak, E., Finley, D., & Varshavsky, A. (1984). The yeast ubiquitin gene: head-to-tail repeats encoding a polyubiquitin precursor protein. *Nature*, 312:663–666.

Rao, S. T. & Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol.*, 76:241–256.

Rechsteiner, M. (1991). Natural substrates of the ubiquitin proteolytic pathway. *Cell*, 66:615–618.

Richmond, T. J. (1984). Solvent accessible surface area and excluded volume in proteins. *J. Mol. Biol.*, 178:63–89.

Shoichet, B. K. & Kuntz, I. D. (1991). Protein docking and complementarity. *J.Mol.Biol.*, 221:327–346.

Silver, E. T., Gwozd, T. J., Ptak, C., Goebl, M., & Ellison, M. J. (1992). A chimeric ubiquitin conjugating enzyme that combines the cell cycle properties of CDC34 (UBC3) and the DNA repair properties of RAD6 (UBC6): implications for the structure, function and evolution of the E2s. *EMBO J.*, 11:3091–3098.

Sokolik, C. W. & Cohen, R. E. (1992). Ubiquitin conjugation to cytochromes *c*. *J. Biol. Chem.*, 267:1067–1071.

Stoddard, B. L. & Koshland, D. E. J. (1992). Prediction of the structure of a receptor-protein complex using a binary docking method. *Nature*, 358:774–776.

Vierstra, R. D., Langan, S. M., & Schaller, G. E. (1986). Complete amino acid sequence of ubiquitin from the higher plant *Avena sativa*. *Biochemistry*, 25:3105–3108.

Vijay-Kumar, S., Bugg, C. E., & Cook, W. J. (1987). Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.*, 194:531–544.

Wilkinson, K. D. (1988). Purification and structural properties of ubiquitin. In Rechsteiner, M., editor, *Ubiquitin*, pages 5–38. Plenum Press.

Wilkinson, K. D., Cox, M. J., O'Connor, L. B., & Shapira, R. (1986). Structure and activities of a variant ubiquitin sequence from bakers' yeast. *Biochemistry*, 25:4999–5004.

Chapter 3

Atomic solvation parameters in the analysis of protein-protein docking results[1]

3.1 Introduction

Although the magnitude of the energetic contributions of solvation and hydrophobic effects to the free energy of interaction of biomolecules is not known precisely, that these effects do make significant contributions to such interactions is undisputed [see, for example (Dill, 1990; Rose & Wolfenden, 1993; Ben-Naim & Mazo, 1993)]. One limitation of many molecular simulations is a lack of consideration of these effects. For molecular simulations that involve many energy calculations and large movements of the molecules involved, such as molecular docking or protein folding simulations, the inclusion of explicit bulk solvent is currently impractical. The development of approximate methods for calculating an energy term representative of the desolvation that occurs during protein folding and biomolecular association has been widely pursued over the last several years [some examples include (Chothia, 1974; Guy, 1985; Eisenberg & McLachlan, 1986; Ooi et al., 1987; Abraham & Leo, 1987; Eisenberg et al., 1989; Still et al., 1990; Wesson & Eisenberg, 1992; Horton & Lewis, 1992; Stouten et al., 1993; Abagyan & Totrov, 1994)].

---

[1] A version of this chapter has been published as: M.D. Cummings, T.N. Hart, & R.J. Read, 1995, *Atomic solvation parameters in the analysis of protein-protein docking results*, Protein Sci. *4*:2087-2099. Reprinted with the permission of Cambridge University Press.

Surveys of protein-protein interfaces show that the average interface tends to be more hydrophobic than the average solvent-exposed protein surface (Chothia & Janin, 1975; Argos, 1988; Janin et al., 1988; Korn & Burnett, 1991). This generalization applies to a variety of protein-protein complexes, although exceptions have been noted (Korn & Burnett, 1991). Two recent methods (Young et al., 1994; Vakser & Aflalo, 1994) and a more detailed study of a specific interface (Clackson & Wells, 1995) also suggest an important role for hydrophobicity in intermolecular interactions. While not conclusive, these observations and developments suggest an important role for hydrophobic interactions in biomolecular recognition.

In a study of thirty-eight protein-protein complexes, Young *et al.* (1994) found that, in two-thirds of the cases, there was significant overlap between the most hydrophobic cluster of surface residues and the ligand binding site. For all but one of the other complexes the hydrophobic cluster showing more than 30% overlap with the ligand binding site was ranked second, third, or fourth most hydrophobic for that protein (in one case the cluster was ranked sixth). This method may be useful in identifying interaction sites on proteins.

Vakser & Aflalo (1994) developed, from their earlier geometric docking algorithm [citations in (Vakser & Aflalo, 1994)] a “hydrophobic docking method” that uses a reduced molecular representation. Non-hydrophobic atoms are omitted from the probe molecule and surface of the target. The relative contribution of hydrophobic interactions to the total intermolecular contact are thus exaggerated. The method also provides for a crude approximation of flexibility (atom deletion), and significantly reduces the computational cost of a docking simulation (fewer atoms). In two of four protein-protein test systems the resolution was very good with the original geometric docking method, and in both of these cases this was improved slightly with the new method. In the other two systems the relatively poor resolution obtained with the original method was significantly improved with the reduced hydrophobic representation.

Clackson & Wells performed an extensive study of the complex of human growth hormone (hGH) and its receptor [hGHR; (Clackson & Wells, 1995)]. Residues were systematically replaced with alanine, and the effect on affinity was measured. For the receptor, replacement of approximately half of the surface buried upon hGH binding had minimal effect on affinity. This included hydrophobic, polar, and charged residues. The central and most hydrophobic part of the interface made the largest contribution to complex stability. Similar results were obtained when hGH residues were replaced. For both proteins, truncation of hydrophobic sidechains had much greater effect on binding than substitution of polar or charged residues.

In 1974 Chothia (1974), extending the work of Kauzmann (1959), and others [cited, most extensively perhaps, in the comprehensive review of Dill (1990)], suggested that surface burial yielded 24 cal Å^{-2} regardless of the chemical nature of the surface. The seminal work of Eisenberg & McLachlan (1986) formalized this concept and developed a simple empirical relationship between the nature and size of the accessible surface of amino acid sidechains and the free energy of transfer of sidechain derivatives from water to octanol. From this relationship they developed a formula for calculating the energetic cost of removing a given protein surface from exposure to water (Eisenberg & McLachlan, 1986;

Eisenberg et al., 1989). Many workers have pursued similar approaches, and this has led to the dissemination of several different atomic solvation parameter (ASP) sets, which vary considerably in magnitude (see the citations at the end of the first paragraph of this chapter). Examples have been reported of the addition of this type of solvation correction to protein folding simulations involving standard (*in vacuo*) potential functions, and the results obtained have generally been promising (Wesson & Eisenberg, 1992; Williams et al., 1992; von Freyberg et al., 1993; Schiffer et al., 1993; Stouten et al., 1993).

Our own work involves the development of molecular docking methods. One of the key issues in the docking problem is the ability to distinguish correct and incorrect dockings on the basis of the calculated interaction energy for the two molecules of interest (here correctness implies similarity to a "known" correct answer). Current methods that sample large configuration spaces must, to achieve their goal in a practical period of time, use relatively crude potential functions that are not always able to distinguish correct dockings in all biological systems [this is discussed in Chapter 2; see, for example (Shoichet & Kuntz, 1991; Bacon & Moult, 1992; Hart & Read, 1992; Cherfils et al., 1994; Totrov & Abagyan, 1994; Cummings et al., 1995)]. This limitation can, at least in some cases, be surmounted by the judicious incorporation of non-energetic information into the experimental design and/or the analysis of simulation results (see preceding citations).

To date our own rigid-body docking simulations have employed a standard Lennard-Jones plus Coulomb potential function, with no explicit consideration of solvent effects [Figure 3.1; (Hart & Read, 1992; Hart & Read, 1994; Cummings et al., 1995)]. Wodak & Janin (1978) first used the accessible surface area (ASA) of the buried surface as a criterion for the ranking of docking results. They concluded that this method was useful for this purpose, although it was not capable, by itself, of unambiguously indicating the best docking. These workers did not dissect the protein-protein interfaces into various chemical types. Recent reports from Janin's group have confirmed that while the total surface area buried upon complex formation is useful in the analysis of docking results, alone it does not allow for the correct ranking of complexes (Cherfils et al., 1991; Cherfils et al., 1994). Shoichet & Kuntz (1991) also reported that neither the total buried surface area nor the solvation free energy, as determined using the ASPs of Eisenberg & McLachlan (1986), was useful in ranking dockings obtained with several different protein-protein systems. Conversely, Horton & Lewis (1992) extended the work of Eisenberg & McLachlan and developed a method for calculating protein-protein interaction energies based solely on hydrogen bonding possibilities and the size and chemical nature of the buried surface. Based on the results presented this method may be quite useful in ranking many protein-protein dockings.

We were interested in incorporating a simple surface-area-based desolvation correction into the energy calculation used in our docking procedure, and were encouraged by some initial results that showed that the application of the correction described by Eisenberg *et al.* (1989) to some of our docking studies with ubiquitin dramatically improved our ability to rank correct dockings as the most energetically favorable [Chapter 2; (Cummings et al., 1995)]. However, the many different ASP sets currently available left us uncertain as to which would be most appropriate for our purposes. We report here the derivation of nine different atomic solvation parameter sets using previously published data. We then compare the final energies calculated for a variety of protein-protein interactions and docking simulations, using our simple Lennard-Jones plus Coulomb potential function, with and without the solvation correction determined from three of our new ASP sets (Figure 3.1).

3.2 Materials and methods

3.2.1 Previously published data

Sidechain transfer free energies for three different systems were taken from Fauchère & Pliska (1983; octanol-water), Radzicka & Wolfenden (1988; cyclohexane-water), and Wolfenden *et al.* (1981; vapor-water). Surface areas for the amino acid sidechain atoms were taken from Eisenberg *et al.* (1989), Wesson & Eisenberg (1992), and Lesser & Rose (1990). Wesson & Eisenberg (1992) did not calculate the sidechain area for Pro. We followed a procedure identical to those authors and calculated this additional data point for their surface area set. Our docking studies involved the following protein structures, which were taken from the Brookhaven Protein Data Bank (Bernstein et al., 1977): *Streptomyces griseus* proteinase B (SGPB) in complex with the third domain of the ovomucoid inhibitor from turkey [OMTKY3; 3sgb; (Read et al., 1983)], native SGPB (Delbaere et al., 1975; Sawyer et al., unpublished coordinates) monoubiquitin [1ubq; (Vijay-Kumar et al., 1987)], and diubiquitin [1aar; (Cook et al., 1992)]. We also used several of the proteinase-inhibitor complexes (PDB codes 2ptc, 1tpa, 2kai, 4cpa, 3cpa, 3sgb, 1cho, and 2tpi), and their respective free energies of association, listed in Table 3.2 of Horton & Lewis (1992).

3.2.2 Calculation of atomic solvation parameters

The atomic solvation parameters were derived by solving a set of simultaneous equations of the form

$$Y = \sum_{i=1}^{5} m_i X_i \tag{3.1}$$

where

Y = the free energy of transfer of the sidechain,

m_i = the ASPs for the five different atom types considered (commonly denoted as $\sigma\Delta$), and

X_i = the ASAs of the various atom types in a particular sidechain.

Multiple linear regressions were performed with the REG module of the SAS/STAT^{TM} software running on an IBM RS6000. The atom types defined in our regressions were C (all carbons), N/O (all uncharged nitrogens and oxygens), N+ (charged nitrogens), O- (charged oxygens), and S (all sulfurs). Given the conditions of the transfer experiments (Fauchère & Pliska, 1983; Radzicka & Wolfenden, 1988; Wolfenden et al., 1981), Glu, Asp, Lys, and Arg were assumed to be completely ionized, and His was assigned a charge of +0.2 (this ratio was not optimized). Previous methods have assigned the charge to the most exposed of the relevant sidechain heteroatoms (Eisenberg & McLachlan, 1986; Eisenberg et al., 1989; Wesson & Eisenberg, 1992; Schiffer et al., 1993; Stouten et al., 1993). This approach seemed rather arbitrary to us, and in the present work we account for resonance distribution of sidechain charges by describing the relevant heteroatoms as linear combinations of two atom types [this approach was taken by Stouten et al. (1993), for His only]. The sidechain Ns of Arg and His are therefore different from that of Lys. For Glu and Asp each sidechain O was described as 50% N/O and 50% O-. For Arg each of the three guanidino Ns was described as 67% N/O and 33% N+. For His each imidazole N was described as 10% N+ and 90 % N/O. These charge distribution ratios were not optimized. All of the data used in the regressions is shown in Tables 1 and 2.

3.2.3 Accessible surface area calculations

All calculations were performed on Silicon Graphics R4000 computers (Indy, Crimson, or Indigo 2XZ). We used the atomic radii of Shrake & Rupley (1973) and a probe radius of 1.4 Å. Surface area calculations were performed with the VADAR program (under development at the University of Alberta; personal communication from D.S. Wishart) which incorporates the ANAREA program (Richmond, 1984). Scatter plots were prepared with the GRAPH module of the program SETOR (Evans, 1993).

The energetic correction for desolvation (E_{desolv}), which was added to the interaction energies obtained with our standard Lennard-Jones plus Coulomb potential function, was calculated according to

$$E_{desolv} = \sum_{i=1}^{5} m_i \Delta A_i \tag{3.2}$$

where

E_{desolv} = the desolvation energy correction,

m_i = the ASPs for the various atom types considered, and

ΔA_i = the changes in ASA of the various atom types that occur upon complex formation.

The charge and surface area assignments described in the preceding section for the ASP derivations were also used in our desolvation calculations.

3.2.4 Docking simulations and energy calculations

Figure 3.1 shows a flowchart of the procedure we used in applying the ASP-based desolvation correction to the analysis of docking results obtained with BOXSEARCH. Our rigid-body multiple start Monte Carlo docking method (Hart & Read, 1992;

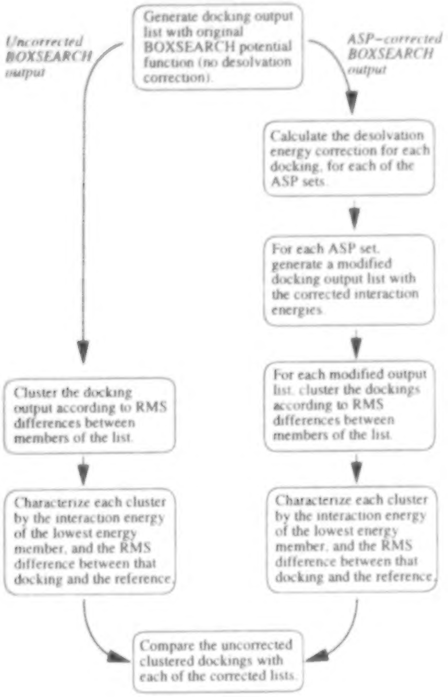

Generate docking output list with original BOXSEARCH potential function (no desolvation correction).
Uncorrected BOXSEARCH output
ASP–corrected BOXSEARCH output
Calculate the desolvation energy correction for each docking, for each of the ASP sets.
For each ASP set, generate a modified docking output list with the corrected interaction energies.
Cluster the docking output according to RMS differences between members of the list.
For each modified output list, cluster the dockings according to RMS differences between members of the list.
Characterize each cluster by the interaction energy of the lowest energy member, and the RMS difference between that docking and the reference.
Characterize each cluster by the interaction energy of the lowest energy member, and the RMS difference between that docking and the reference.
Compare the uncorrected clustered dockings with each of the corrected lists.

Figure 3.1: Flowchart showing the procedure for re-analysis of docking results. The effects of incorporation of various ASP sets into the analysis of docking results obtained with the BOXSEARCH program were studied, to determine which ASP set was most compatible with the simple Lennard-Jones plus Coulomb potential function used in BOXSEARCH.

Hart & Read, 1994), and methods of protein structure preparation [Chapter 2; (Cummings et al., 1995)], have been described in detail elsewhere. Since the present work comprises a re-analysis of docking results (see Figure 3.1), as well as some simple protein-protein interactions, our docking method is not discussed here. We minimized the series of proteinase-inhibitor complexes with a rigid-body conjugate gradient minimizer (unpublished program of Trevor Hart). Polar hydrogen positions were not optimized for this series of complexes. These differences account for the slight increase in calculated interaction energy for the 3sgb complex listed in Figure 3.9, over that listed in Table 3.5. Manipulation of protein structures was performed with INSIGHTII (Biosym Technologies, San Diego). The interaction energies of some previously reported dockings (Hart & Read, 1992) have been recalculated with an updated version of the potential function used in our energy calculations (Table 2.5). For all cases studied the energy calculated with our unmodified (no solvation term) potential function is shown as a reference. The potential function used in BOXSEARCH is denoted as

$$E_{uncorr} = E_{vdW} + E_{elec} \tag{3.3}$$

where

E_{uncorr} = the uncorrected interaction energy,
E_{vdW} = the van der Waals energy contribution, and
E_{elec} = the electrostatic energy contribution.

For the present work the docking simulations were performed with the unmodified BOXSEARCH potential function [equation (3.3); for the explicit form, see (Hart & Read, 1992)] and Appendix A]. However, the final energy of each accepted docking was also (see Figure 3.1) calculated according to

$$E_{corr} = E_{vdW} + E_{elec} + E_{desolv} \tag{3.4}$$

where

E_{corr} = the corrected interaction energy, and
E_{desolv} = the desolvation energy contribution calculated according to equation (3.2).

For the series of simple protein-protein associations that we analyzed we calculated the interaction energy in the same way with equations (3.3) and (3.4).

		area of atom type ($Å^2$)				
area set	residue	C	N/O	N+	O-	S
Eisenberg et al. (1989)	Ala	60.0	0	0	0	0
	Arg	81.0	80.7	40.3	0	0
	Asn	41.0	73.0	0	0	0
	Asp	51.0	28.0	0	28.0	0
	Cys	25.0	0	0	0	27.0
	Gln	59.0	83.0	0	0	0
	Glu	82.0	31.5	0	31.5	0
	His	120.0	43.2	4.8	0	0
	Ile	153.0	0	0	0	0
	Leu	151.0	0	0	0	0
	Lys	114.0	0	58.0	0	0
	Met	126.0	0	0	0	34.0
	Phe	184.0	0	0	0	0
	Pro	109.0	0	0	0	0
	Ser	41.0	34.0	0	0	0
	Thr	80.0	26.0	0	0	0
	Trp	204.0	26.0	0	0	0
	Tyr	152.0	39.0	0	0	0
	Val	125.0	0	0	0	0
Wesson & Eisenberg (1992)	Ala	137.0	0	0	0	0
	Arg	154.0	82.0	41.0	0	0
	Asn	98.0	96.0	0	0	0
	Asp	111.0	37.5	0	37.5	0
	Cys	91.0	0	0	0	79.0
	Gln	130.0	91.0	0	0	0
	Glu	143.0	34.5	0	34.5	0
	His	182.0	47.7	5.3	0	0
	Ile	226.0	0	0	0	0
	Leu	221.0	0	0	0	0
	Lys	185.0	0	60.0	0	0
	Met	191.0	0	0	0	40.0
	Phe	260.0	0	0	0	0
	Pro[a]	185.0	0	0	0	0
	Ser	105.0	43.0	0	0	0
	Thr	140.0	40.0	0	0	0
	Trp	279.0	26.0	0	0	0
	Tyr	233.0	38.0	0	0	0
	Val	196.0	0	0	0	0
Lesser & Rose (1990)	Ala	71.9	0	0	0	0
	Arg	81.6	90.2	45.1	0	0
	Asn	35.1	90.2	0	0	0
	Asp	38.6	39.8	0	39.8	0
	Cys	38.5	0	0	0	65.0
	Gln	61.2	94.2	0	0	0
	Glu	65.5	41.4		41.4	0
	His	120.2	37.8	4.2	0	0
	Ile	150.1	0	0	0	0
	Leu	157.8	0	0	0	0
	Lys	112.5	0	74.6	0	0
	Met	128.4	0	0	0	36.4
	Phe	184.4	0	0	0	0
	Pro	111.0	0	0	0	0
	Ser	44.3	41.5	0	0	0
	Thr	81.1	33.5	0	0	0
	Trp	199.1	29.8	0	0	0
	Tyr	145.2	52.9	0	0	0
	Val	128.4	0	0	0	0

Table 3.1: Three sets of surface area data for regression analysis. [a]The surface area for this residue was calculated by us, as described in the Methods section.

3.3 Results and Discussion

3.3.1 Development of new ASPs

We chose the surface area sets used in the development of two frequently encountered ASP sets (Eisenberg et al., 1989; Wesson & Eisenberg, 1992) as well as the much more comprehensive set determined by Lesser & Rose (199). Table 3.1 shows the total areas reported by these groups after re-analysis according to the charge and area assignment scheme described above. The main difference between our method and those reported earlier [see, for example, (Eisenberg & McLachlan, 1986; Eisenberg et al., 1989; Wesson & Eisenberg, 1992; Stouten et al., 1993)] is that, instead of arbitrarily assigning charges to the most exposed heteroatom of a charged sidechain, we distribute the charge evenly over all possible heteroatoms. This is accomplished by describing the relevant sidechain heteroatoms of Glu, Asp, Arg, and His as linear combinations of two atom types.

The surface area data of Eisenberg et al. (1989) and of Lesser & Rose (1990) are fairly similar, whereas the data of Wesson & Eisenberg (1992) differ considerably (Table 3.1). This is especially pronounced for the carbon atoms, and reflects the different methods used in the surface area calculations. Wesson & Eisenberg used isolated *sidechains* from four protein structures. Surface areas were calculated for a number of copies (twenty of each in most cases) of the residue of interest, in the absence of the remainder of the protein structure. For the other surface area sets shown in Table 3.1, residues in extended Gly-X-Gly sequences were used in the surface area calculations. The method of Wesson & Eisenberg results in greater exposure of sidechains atoms, especially C_βs. The surface areas derived by this method are appropriate for correlation to the cyclohexane-water and vapor-water transfer energies, since these experimental values were obtained with *sidechain* analogs. On the other hand, we consider the comprehensive set of average sidechain surface areas of Lesser & Rose (1990) to be the most appropriate for the development of octanol-water-based ASPs, since these transfer energy data were obtained with blocked amino acid derivatives. These authors calculated surface areas for several hundred copies of each amino acid (Trp was the minimum at 157, Gly the maximum at 1004) in extended Gly-X-Gly tripeptides. 10,937 residues from 61 protein structures were used in their analysis. These data should provide the best "average" conformation for each of the sidechains. The similarity between these conformations and those adopted by the relevant small molecules in vapor, cyclohexane, or octanol is not considered here. However, our results with regressions of different pairings of transfer energy and surface area data sets show that small changes in surface areas do not have major effects on the derived ASPs (Tables 3.1, 3.2, and 3.4). The different atom classification schemes and atomic radii used in these previous studies also account, in part, for the different areas obtained.

In conjunction with the above-described surface area data, we chose frequently-cited transfer energy data for amino acid analogs studied in three different solvent systems (Table 3.2). The ASPs originally derived by Eisenberg and co-workers (Eisenberg & McLachlan, 1986; Eisenberg et al., 1989) were based on the octanol-water partition data of Fauchère & Pliska (1983), which is possibly the most commonly used set of amino acid transfer energy data. Since we are approximating protein *desolvation*, one might expect vapor-water transfer to provide the most accurate model. Indeed, the vapor-water data of Wolfenden et al. (1981) have also been used extensively, and increasingly, for this purpose. One possible limitation of this data set as the basis for ASP derivation, recently noted by Schiffer et al. (1993), is the fact that this data set is a compilation of results from several different laboratories [Schiffer et al. (1993), and refs 30-35 therein]. Wolfenden's group has also reported the transfer energies for amino acid sidechains in the cyclohexane-water system (Radzicka & Wolfenden, 1988). These data are also compiled from several

	transfer free energy[a] (kcal mol^{-1})		
residue	Fauchère	Radzicka	Wolfenden
Ala	0.42	1.81	1.94
Arg	-1.37	-14.92	-19.92
Asn	-0.82	-6.64	-9.68
Asp	-1.05	-8.72	-10.95
Cys	1.34	1.28	-1.24
Gln	-0.30	-5.54	-9.38
Glu	-0.87	-6.81	-10.24
His	0.18	-4.66	-10.27
Ile	2.46	4.92	2.15
Leu	2.32	4.92	2.28
Lys	-1.35	-5.55	-9.52
Met	1.68	2.35	-1.48
Phe	2.44	2.98	-0.76
Pro	0.98	–	–
Ser	-0.05	-3.40	-5.06
Thr	0.35	-2.57	-4.88
Trp	3.07	2.33	-5.88
Tyr	1.31	-0.14	-6.11
Val	1.66	4.04	1.99

Table 3.2: Three sets of transfer free energies for regression analysis. [a]Transfer free energies from Fauchère & Pliska (1983; Fauchère), Radzicka & Wolfenden (1988; Radzicka), and Wolfenden et al. (1981; Wolfenden).

different sources.

Given the low water content of cyclohexane (compared to octanol), as well as the absence of hydrogen-bonding groups, partitioning into cyclohexane should provide a “purer” measure of the relative susceptibilities of the various sidechains to non-specific dispersion forces only (Radzicka & Wolfenden, 1988). Relevant to this argument is work from the laboratories of Wolfenden (Radzicka et al., 1993) and Testa (Tsai et al., 1993; Fan et al., 1994). Solutes were shown to “drag” water from the aqueous to the organic phase of biphasic partition systems (Tsai et al., 1993; Fan et al., 1994). In octanol, this effect is further complicated by the high water content of this solvent (Fan et al., 1994). Unfortunately, amino acid derivatives were not specifically investigated in these studies, although there appears to be no reason to expect significantly different results with such solutes. In a related study, Wolfenden's group (Radzicka et al., 1993) found that, with several different sidechain and backbone analogs, such entrainment of water into (the relatively non-polar, and dry, solvent) cyclohexane was minimal. Such complications cloud the meaning of partition data, and serve to reinforce the empirical nature of ASP development.

Ben-Naim (1994) and others [for example, see (Holtzer, 1994)] have questioned the volume correction used to derive one of the commonly encountered ASP sets (Sharp et al., 1991), and shown that this correction is unnecessary. We have avoided any such corrections in the present work.

Unlike the surface area data sets, choosing between the transfer energies derived from the different solvent systems is not straightforward. The correlation coefficients of all nine multiple linear regressions and the root-mean-square differences between

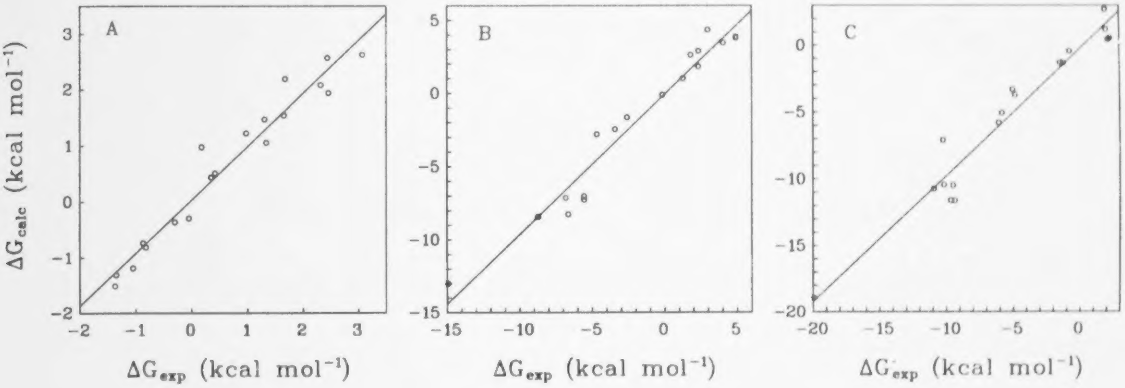


Figure 3.2: Correlation of calculated and experimental transfer free energies of amino acid analogs for 3 different ASP sets. **A:** transfer of blocked amino acids in the octanol-water system (ASP set 3); **B:** transfer of sidechain analogs in the cyclohexane-water system (ASP set 5); **C:** transfer of sidechain analogs in the vapor-water system (ASP set 8).

area[b]	transfer free energy data[c]					
	Fauchère		Radzicka		Wolfenden	
	$\mathcal{R}$[d]	rms[e]	$\mathcal{R}$	rms	$\mathcal{R}$	rms
Eisenberg	.90	.36(8.2)	.93	1.2(5.8)	.91	1.4(6.4)
Wesson	.93	.30(6.8)	**.94**	**1.1(5.5)**	**.92**	**1.4(5.9)**
Lesser	**.93**	**.30(6.8)**	.90	1.4(7.0)	.88	1.7(7.4)

Table 3.3: Statistics for various transfer free energy - surface area regressions. [a]Statistics are from the multiple linear regressions as described in the Methods section. ASP sets 3, 5, and 8 are in boldface. [b]Fits were obtained with the surface areas reported in Eisenberg et al. (1989; Eisenberg), Wesson & Eisenberg (1992; Wesson), and Lesser & Rose (1990; Lesser). [c]Transfer free energies from Fauchère & Pliska (1983; Fauchère), Radzicka & Wolfenden (1988; Radzicka), and Wolfenden et al. (1981; Wolfenden). [d]The corrected (for the number of parameters in the regression) multiple correlation coefficient of that regression. [e]Root-mean-square difference between the experimental and calculated energy for that particular fit. Numbers in parentheses are the rms error represented as a percentage of the total range of the experimental transfer free energy for that particular fit.

the experimental and calculated transfer energies of the sidechain derivatives are all very good (Table 3.3 and Figure 3.2). The ASP sets cannot be distinguished on either of these bases. This is somewhat surprising considering the differences in the surface areas noted above. All three ASP sets derived using Fauchère's (1983) octanol-water data are similar to each other, as well as to Eisenberg's octanol-water ASP sets (Eisenberg & McLachlan, 1986; Eisenberg et al., 1989). Schiffer et al. (1993) recently reported much lower charged ASPs based on octanol-water transfer energies of eight Ala-X-Ala tripeptides (their C and N/O ASPs were similar to ours). The biggest variation in our three octanol-water ASP sets is in the charged atom parameters (Table 3.4). ASPs derived from cyclohexane-water transfer energies have C and S ASPs similar to the octanol-water sets while the other heteroatom parameters are closer to those of the vapor-water sets (Table 3.4). The ASP sets derived from Wolfenden's vapor-water data (Wolfenden et al., 1981) all have negative carbon parameters (ASP sets 7-9 in Table 3.4). This casts doubt on the utility of the vapor-water data, because we generally expect the burial of carbon atoms to be energetically

			atomic solvation parameters[c] (cal $Å^{-2}$ mol^{-1})				
ASP set #	ΔG^{a}_{exp}	area[b]	$\Delta\sigma(C)$	$\Delta\sigma(N/O)$	$\Delta\sigma(N^{+})$	$\Delta\sigma(O^{-})$	$\Delta\sigma(S)$
1	Fauchère	Eisenberg	15±2	-12±4	-44±7	-44±12	20±12
2	Fauchère	Wesson	16±2	-8±3	-43±5	-30±8	16±5
3	**Fauchère**	**Lesser**	**18±2**	**-7±3**	**-34±4**	**-20±8**	**18±6**
4	Radzicka	Eisenberg	22±8	-117±14	-165±22	-209±39	-14±40
5	**Radzicka**	**Wesson**	**14±8**	**-107±13**	**-172±20**	**-177±32**	**-12±21**
6	Radzicka	Lesser	22±11	-100±16	-133±22	-129±36	-5±30
7	Wolfenden	Eisenberg	-12±10	-166±17	-187±27	-236±48	-86±49
8	**Wolfenden**	**Wesson**	**-25±10**	**-159±16**	**-199±25**	**-217±38**	**-66±25**
9	Wolfenden	Lesser	-19±13	-150±19	-156±26	-167±43	-60±36

Table 3.4: Atomic solvation parameters derived from various regressions. Atomic solvation parameters were derived from the various data sets by multiple linear regression as described in the Methods section. [a]Transfer free energies from Fauchère & Pliska (1983; Fauchère), Radzicka & Wolfenden (1988; Radzicka), and Wolfenden et al. (1981; Wolfenden). [b]Fits were obtained with the surface areas reported in Eisenberg et al. (1989; Eisenberg), Wesson & Eisenberg (1992; Wesson), and Lesser & Rose (1990; Lesser). [c]Parameters are shown ± their standard error.

favorable (a negative ASP implies that exposure of that atom type to solvent is more energetically favorable).

Unfortunately, for the reasons noted above, we cannot use one surface area data set in regressions with the three different transfer energy data sets. Indeed, some of the pairings of surface area and transfer energies in the regressions shown in Tables 3 and 4 could not be justified. On the other hand, they do serve to underline the point that small changes in surface areas will not have a large effect on the ASPs derived from those areas. Based upon the match between the molecules used for the transfer energy experiments and for the surface area calculations, the most meaningful regressions are those represented by ASP sets 3, 5, and 8 (Table 3.4). For this reason, and also to limit the amount of data to be analyzed, we chose to evaluate only three ASP sets. This allowed us to observe the differences in the energy corrections obtained using ASPs derived from transfer energies determined with the three different solvent systems.

Due to methodological differences it was difficult for us to compare several previously published reports concerning the development and evaluation of ASPs (Eisenberg & McLachlan, 1986; Ooi et al., 1987; Eisenberg et al., 1989; Wesson & Eisenberg, 1992;
Horton & Lewis, 1992; Schiffer et al., 1993; Stouten et al., 1993). Our purpose in the current work was to re-evaluate several extensively cited sets of experimental data in a consistent manner and then empirically evaluate the appropriateness of the ASP sets thus derived as complements to the Lennard-Jones plus Coulomb potential function used in our docking simulations. A related study has recently been reported, wherein three published ASP sets (and one new set based on different octanol-water transfer data) were evaluated as complements to the AMBER potential function in a protein folding/molecular dynamics study (Schiffer et al., 1993).

3.3.2 Evaluation of ASPs: criteria in docking

For the analysis of protein-protein docking results, two energy differences are of primary concern. The first is the energy difference between a reference (correct) docking and the lowest energy truly incorrect docking. We assume that correct dockings of lower energy than the lowest energy incorrect docking would be identifiable in the absence of a reference docking. In previous work we have used a difference of 2 Å RMS as the cutoff for similarity between dockings [Chapter 2; (Hart & Read, 1992; Hart & Read, 1994; Cummings et al., 1995); see also Chapters 4 - 6], although some of our minimization studies suggest that this may be overly restrictive (results not shown). The second is the energy difference between the lowest energy correct docking and the lowest energy clearly incorrect docking. This second energy difference is probably the more important of the two, since in a truly predictive situation a reference docking (a complex configuration *known* to be correct) is not

			modified docking results								
			ASP set 3			ASP set 5			ASP set 8		
rank^b	RMS^c (Å)	E^d_{uncorr} (kcal/mol)	rank	RMS (Å)	E^d_{corr} (kcal/mol)	rank	RMS (Å)	E_{corr} (kcal/mol)	rank	RMS (Å)	E_{corr} (kcal/mol)
Complexed SGPB-OMTKY3 (see Figure 3.3)											
ref	0.0	-98.5	ref	0.0	-111.4	ref	0.0	-69.1	ref	0.0	-15.1
1	0.4	-93.4 (5.1)	1	0.4	-102.2 (9.2)	1	0.5	-47.4 (22.7)	1	0.5	0.8 (15.9)
4	16.9	-55.8 (37.6)	4	16.9	-59.4 (42.3)	5	19.0	2.6 (50.0)	5	16.9	32.0 (31.2)
Native SGPB-OMTKY3 (see Figure 3.4)											
ref	0.0	-80.4	ref	0.0	-93.4	ref	0.0	-54.2	ref	0.0	-2.1
1	1.8	-74.4 (6.0)	1	1.8	-82.7 (10.7)	1	1.8	-33.5 (20.7)	1	1.8	11.5 (13.6)
3	17.9	-55.0 (19.4)	3	11.7	-63.2 (19.5)	3	11.7	-20.1 (13.4)	2	20.7	13.4 (1.9)
Complexed SGPB-FRAG1 (see Figure 3.5)											
ref	0.0	-41.9	ref	0.0	-49.9	ref	0.0	-35.3	ref	0.0	-9.4
1	1.1	-39.1 (2.8)	1	1.1	-45.6 (4.3)	1	1.1	-25.7 (9.6)	3	1.1	-1.7 (7.7)
2	7.1	-33.4 (9.6)	2	5.6	-35.1 (10.5)	2	21.9	-16.1 (9.6)	1	20.8	-3.4 (-1.7)
Complexed Diubiquitin (see Figure 3.6)											
ref	0.0	-79.2	ref	0.0	-90.8	ref	0.0	-29.8	ref	0.0	32.6
1	0.4	-76.6 (2.6)	1	0.4	-83.2 (7.6)	2	0.6	-12.4 (17.4)	69	0.6	43.3 (10.7)
4	15.5	-49.5 (27.1)	5	13.7	-48.9 (34.3)	1	5.3	-19.4 (-7.0)	1	5.3	13.1 (-30.2)
Modified Diubiquitin (see Figure 3.7)											
ref	0.0	-68.6	ref	0.0	-79.8	ref	0.0	-27.4	ref	0.0	29.0
1	0.5	-65.5 (3.1)	1	0.5	-72.5 (7.3)	1	0.6	-30.9 (-3.5)	82	0.6	42.7 (13.7)
2	17.4	-51.1 (14.4)	4	17.4	-54.7 (17.8)	2	19.8	-27.6 (3.3)	1	12.0	10.4 (-32.3)
Modified Monoubiquitin (see Figure 3.8)											
ref	0.0	-39.0	ref	0.0	-48.3	ref	0.0	-20.6	ref	0.0	16.1
43	1.6	-37.0 (2.0)	25	1.6	-42.4 (5.9)	7	1.6	-0.9 (19.5)	16	1.6	34.0 (17.9)
1	4.6	-52.1 (-15.1)	1	4.6	-56.6 (-14.2)	1	4.2	-9.2 (-8.3)	1	4.2	21.7 (-12.4)
2	23.1	-51.3 (-14.3)	2	3.3	-54.9 (-12.5)	2	2.4	-8.5 (-7.6)	2	2.4	23.6 (-10.5)
3	3.3	-49.0 (-12.0)	3	23.1	-51.4 (-9.0)	3	3.9	-7.0 (-6.1)	3	21.7	23.7 (-10.4)
						4	3.3	-5.9 (-5.0)			
						5	16.0	-3.5 (-2.6)			

Table 3.5: Low Energy Dockings and Reference Configurations for the Docking Systems. ^{a}The first line of each section describes the reference configuration for the four rankings of each docking system. The second line of each section represents the lowest energy clearly correct (within 2 Å RMS of the relevant reference configuration) docking for that ranking. The third line of each section represents the lowest energy clearly incorrect docking (further than 5 Å RMS from the relevant reference) for that ranking. For modified (mono)ubiquitin a few more of the interesting low energy clusters are also listed (lines 3-7). ^{b}This number represents the ranking of that cluster (1 is the lowest energy). Ref is the reference configuration. ^{c}Root-mean-square difference (all-atoms) between that docking and its reference. ^{d}E_{uncorr} is the uncorrected (for desolvation) BOXSEARCH interaction energy. E_{corr} is the corrected (for desolvation, according to the indicated ASP set) BOXSEARCH interaction energy. The numbers in parentheses on line 2 of each section represent the energy difference between the lowest energy clearly correct docking (line 2) and the reference (line 1) configuration (*i.e.* E_{line2} - E_{line1}). A negative value for this energy difference indicates that this docking had a more favorable (lower) interaction energy than the reference configuration. The numbers in parentheses on line 3 (or below) of each section represent the energy difference between that docking (line 3, or lines 3-7 in the case of modified monoubiquitin) and the lowest energy clearly correct docking (line 2) for that ranking (*i.e.* E_{line3} - E_{line2}). A negative value for this energy difference indicates that the docking (in lines 3-7) had a more favorable interaction energy than the lowest energy correct docking. Summing the two numbers in parentheses gives the energy difference between the docking (lines 3-7) and the reference (line 1; *i.e.* E_{line3} - E_{line1}). These critical energy differences are discussed in the text.

available. In Table 3.5 we summarize the information relevant to these two energy differences, for the six docking systems studied here.

With the above considerations in mind, we re-analyzed docking results from several systems that we had studied previously (see Figure 3.1). For the purpose of the present discussion we consider dockings within 2 Å RMS of the reference configuration to be correct, and dockings 2-5 Å RMS distant from the reference to be close to correct. Dockings further than 5 Å RMS from the reference are considered to be clearly incorrect. In the absence of a known reference structure the definition of correct and close to correct depend on the distance constraints available, as well as the nature of the complex being studied. For all of the systems described in the present work a reference configuration was available to us at the time the docking simulations were performed.

3.3.3 Evaluation of ASPs: docking with SGPB-OMTKY3

In the original description of BOXSEARCH, complexed OMTKY3 and various fragments of OMTKY3 were docked to both native and complexed SGPB (Hart & Read, 1992). Figure 3.3 summarizes results obtained with the complexed forms of both SGPB and OMTKY3. The difference between the interaction energies of the lowest energy correct docking and each cluster is shown for the uncorrected BOXSEARCH potential function (Figure 3.3A), as well as for the desolvation-corrected rankings obtained with ASP sets 3, 5, and 8 (Figures 3.3B, 3.3C, and 3.3D, respectively). From an examination of the energy differences, we see that ASP sets 3 and 5 improve the discrimination between correct and incorrect dockings slightly, while ASP set 8 makes it worse (Figure 3.3, Table 3.5).

Overall, the re-analysis of the docking results obtained with native SGPB and complexed OMTKY3 gave similar results to those observed with complexed SGPB (Figure 3.4). With the exception of ASP set 8, all rankings had correct dockings

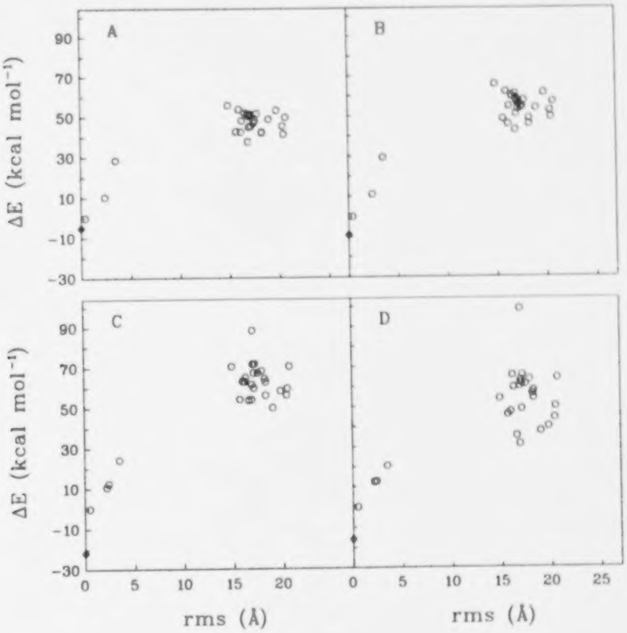


Figure 3.3: Energy differences for complexed SGPB-OMTKY3. The difference between the interaction energy of the lowest energy correct docking (listed in Table 3.5) and each docking in the clustered output list (see Figure 3.1) is shown, as a function of RMS distance from the reference configuration (open circles). The energy difference between the reference configuration and the lowest energy correct docking is also shown (filled diamond with rms = 0.0). **A:** interaction energy calculated with the uncorrected BOXSEARCH potential function; **B:** BOXSEARCH potential function corrected with ASP set 3; **C:** BOXSEARCH potential function corrected with ASP set 5; **D:** BOXSEARCH potential function corrected with ASP set 8.

for the two lowest energy clusters. Both ASP sets 5 and 8 decreased our ability to discriminate between correct and incorrect dockings on the basis of interaction energies (Table 3.5). For ASP set 3, the discrimination of the single lowest energy incorrect docking was not improved (Table 3.5), but the discrimination was improved for the bulk of the incorrect dockings (Figure 3.4). BOXSEARCH corrected with ASP set 8 (Figure 3.4D) again clearly gave the worst ranking.

The differences between the desolvation corrections obtained with the various ASP sets were more apparent with the OMTKY3 reactive site tripeptide [FRAG1 here and

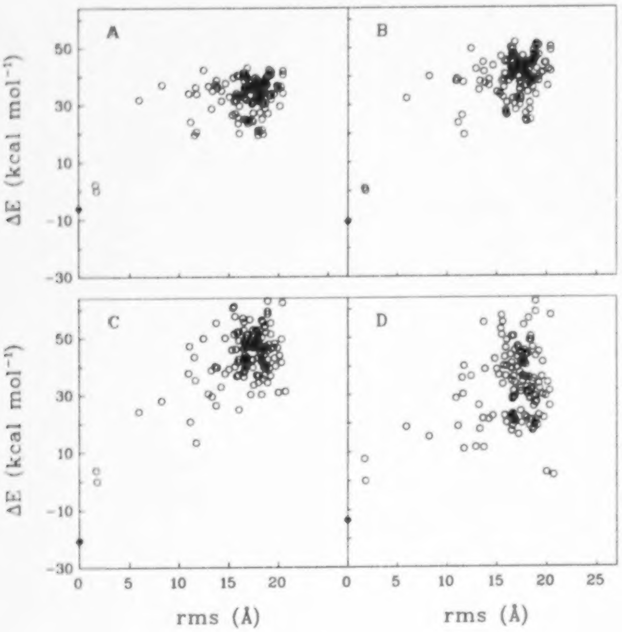


Figure 3.4: Energy differences for native SGPB-OMTKY3. See legend to Figure 3.3.

in Hart & Read (1992)] than with either of the systems described above (compare Figure 3.5 with Figures 3.3 and 3.4). This may reflect the fact that in this case the uncorrected ranking is much less clear than in the two previously discussed SGPB-OMTKY3 systems. All rankings except ASP set 8 had a correct docking for the lowest energy cluster (Table 3.5). Correction with ASP set 5 provided the largest difference between the reference docking and the lowest energy incorrect docking (Figure 3.5C, Table 3.5), whereas ASP set 3 was best at distinguishing between the lowest energy correct and incorrect dockings (Figure 3.5B, Table 3.5). With ASP set 8 the low energy correct docking was energetically indistinct from many of the incorrect dockings (Figure 3.5D, Table 3.5).

One general observation warrants discussion at this point. In the three systems studied so far, we see a tendency for both ASP sets 5 and 8 to add more noise than signal to the rankings. This trend can be intuitively grasped by visually comparing

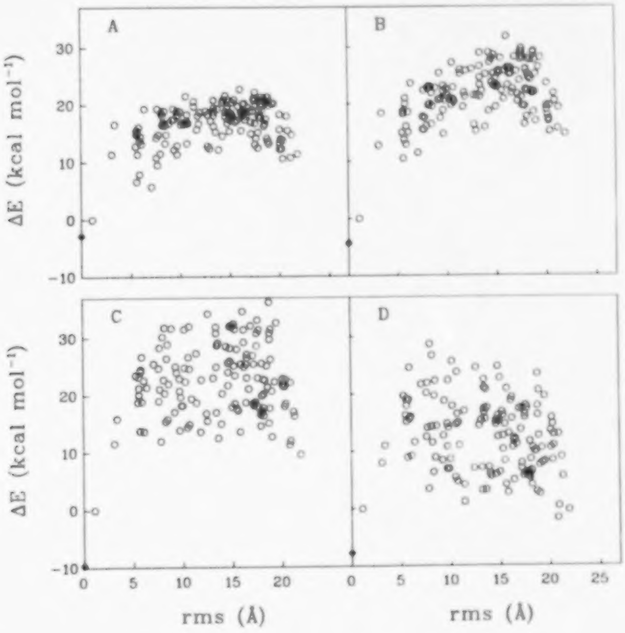


Figure 3.5: Energy differences for complexed SGPB-FRAG1. See legend to Figure 3.3.

plots C and D with A and B in Figures 3.3, 3.4, and 3.5. Many incorrect dockings become (apparently) more energetically favorable, and this increase in favorability (decrease in interaction energy) is relatively greater for the incorrect dockings than for the correct dockings. ASP set 3 adds a smaller correction, which is more often than not in the correct direction. This tendency persists in the docking systems discussed below.

3.3.4 Evaluation of ASPs: docking with ubiquitin

We recently reported the results of docking simulations involving the ubiquitin monomer and dimer [Chapter 2; (Cummings et al., 1995)]. In one set of experiments, we found Eisenberg's ASP-based desolvation correction (Eisenberg et al., 1989) to provide a good complement to our BOXSEARCH interaction energy values; the relative ranking of correct and incorrect dockings was substantially improved. Here we re-analyze the results of docking simulations with three different ubiquitin-ubiquitin pairs, using our new ASP sets.

First we consider docking together the two halves of the diubiquitin structure, the results of which are summarized in Figure 6. The uncorrected docking results are very good, with four of the five lowest energy clusters being correct, or very close to correct, and the energy differences between correct and incorrect dockings show good discrimination (Table 3.5, Figure 3.6A). ASP set 3 marginally improved our ability to distinguish between low energy correct and incorrect dockings, and five of the six lowest energy clusters are now correct, or close to correct (Figure 3.6B). ASP sets 5 and 8 both perform very poorly in this system (Figures 3.6C and 3.6D). With ASP set 8 the ranking of the reference docking is actually inverted relative to the twenty-four lowest energy clusters, and the lowest energy correct docking is ranked 69 (in a

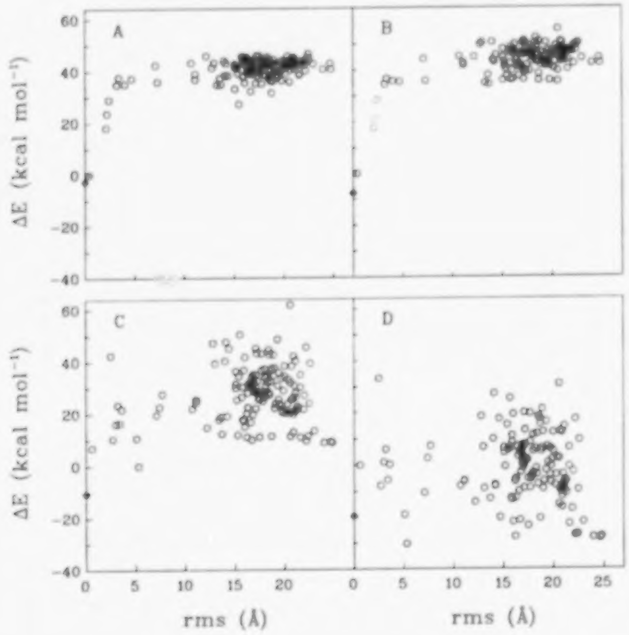


Figure 3.6: Energy differences for complexed diubiquitin. See legend to Figure 3.3.

total of 132; Figure 3.6D, Table 3.5). With ASP set 5 the low energy clusters that are correct, or close to correct, become energetically indistinct from many incorrect clusters (Figure 3.6C). The crucial energy differences are greatly reduced (from 30-40 to 10 kcal mol^{-1}; Figure 3.6, Table 3.5).

We also docked two copies of one half of the diubiquitin molecule to each other (Figure 3.7). In this case BOXSEARCH alone ranks three of the four lowest energy clusters as correct or close to correct, with the lowest energy incorrect docking being cluster 2 (Figure 3.7A, Table 3.5). Correction with ASP set 3 improves this to four of the five lowest energy clusters being correct or close to correct, and the lowest energy incorrect docking is moved to cluster 4 (Figure 3.7B, Table 3.5). The overall discrimination between correct and incorrect dockings is improved. Conversely, ASP sets 5 and 8 significantly disturb the clustering of the low energy dockings, and the crucial energy differences are decreased or inverted (Figure 3.7, Table 3.5).

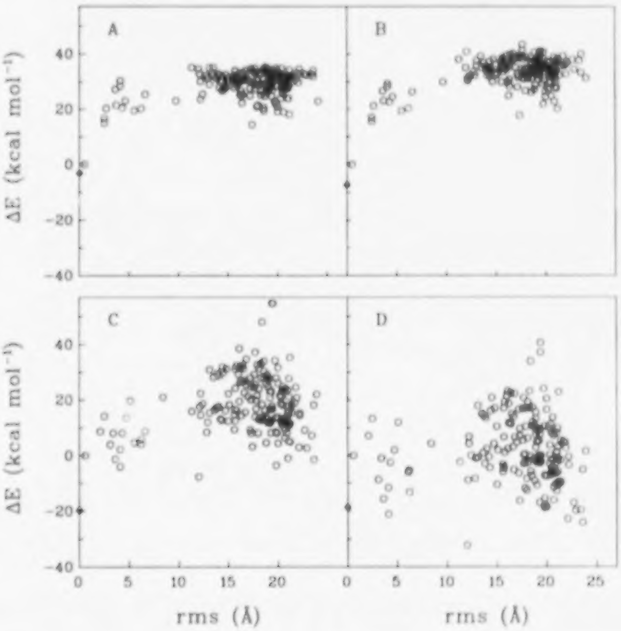


Figure 3.7: Energy differences for modified diubiquitin. See legend to Figure 3.3.

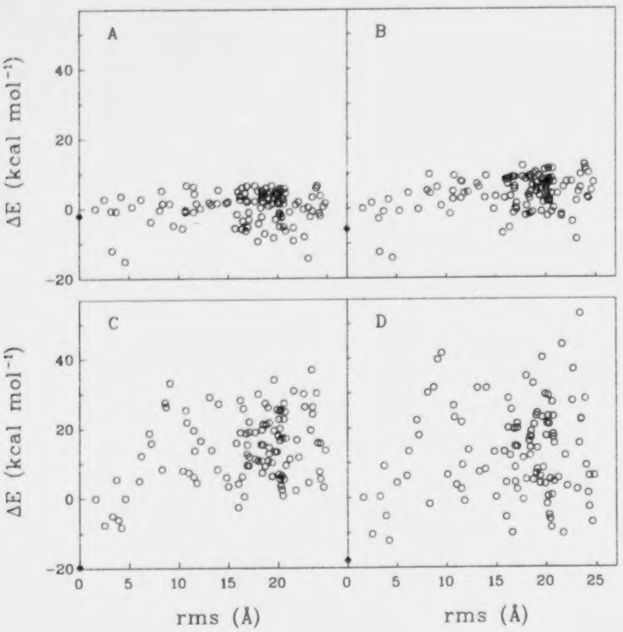


Figure 3.8: Energy differences for native (mono)ubiquitin. See legend to Figure 3.3.

Finally, we re-analyze the docking results that were corrected with the ASPs of Eisenberg *et al.* (1989) in our earlier work [Chapter 2; (Cummings et al., 1995)]. This simulation involved two copies of the ubiquitin monomer in which a crucial Arg residue had been truncated to Ala to facilitate formation of a diubiquitin-like dimer. In this simulation the energy differences between correct and incorrect low energy dockings are not as great as in the other two ubiquitin systems discussed above.

In the uncorrected ranking forty-two clusters are represented by dockings of lower energy than the lowest energy correct docking (Figure 3.8A, Table 3.5). Two of the three lowest energy dockings are close to correct (Table 3.5). ASP set 3 slightly improves the ranking and clustering of the low energy dockings, such that only twenty-four clusters are of lower energy than the lowest energy correct docking (Figure 3.8B, Table 3.5). Also, the two lowest energy clusters are now both close to correct. In contrast to the results discussed above, in this system ASP sets 5 and 8 perform better than ASP set 3 (Figure 3.8, Table 3.5). Correction with ASP set 5 results in a ranking in which only six clusters are of lower energy than the lowest energy correct docking, and six of the seven lowest energy clusters are close to correct, or correct (Figure 3.8C, Table 3.5). Application of ASP set 8 yields a ranking with fifteen clusters of lower energy than the lowest energy correct docking (Figure 3.8D, Table 3.5). Several of the low energy clusters are close to correct. For both ASP sets 5 and 8 the reference configuration is of lower energy than any of the dockings, in contrast to the ASP set 3 and uncorrected rankings (Table 3.5).

In our original application of this method we applied a set of ASPs similar to ASP set 3 (Eisenberg et al., 1989) to this last ubiquitin system. This improved the relative ranking of the lowest energy clusters in the same way that ASP set 3 did in the present work (Figure 3.8, Table 3.5). At that time we noted that, while the energetic distinction was still small, the chemical nature of the interface of cluster 2 in the uncorrected ranking (Table 3.5) was markedly different than our other low energy (and closer to correct) dockings. This interface was also quite different from the averages reported for a survey of subunit-subunit interfaces (Janin et al., 1988). The highly polar and charged nature of the interface of this incorrect docking explains, at least in one case, why ASP sets 5 and 8 perform better than set 3 in this docking system.

3.3.5 Evaluation of ASPs: a series of protease-inhibitor complexes

In protein-protein docking our primary interest is in the use of the BOXSEARCH potential function to discriminate between correct and incorrect dockings of the same complex. However, another way to evaluate the ASPs described above is to study a series of different protein-protein complexes, as did Horton & Lewis (1992). This is reminiscent of a drug-design scenario wherein one is concerned not only with evaluating dockings of the same ligand, but also with the low energy dockings of many different ligands. Success in this arena requires the ability to rank correctly the favorability of a series of dockings of *the same* ligand, as well as the favorability of a series of *different* ligands. We took several of the protease-inhibitor complexes studied by Horton & Lewis (1992) and calculated the interaction energies of these complexes with the uncorrected BOXSEARCH potential function, as well as with the corrected potential function, using ASP sets 3, 5, and 8.

Figure 3.9A shows the correlation of the experimental free energies of interaction with the interaction energies calculated by the BOXSEARCH potential function alone. Although our potential function has difficulty with the ranking of complexes with similar interaction free energies, it is clear that our calculated interaction energies are strongly correlated to the experimental free energies of interaction (Figure 3.9A). The calculated interaction energies are much larger negative values and span a range of 75 kcal mol^{-1}, whereas the experimental free energies span the narrower range of 13 kcal mol^{-1}.

Correction of the BOXSEARCH potential function with the octanol-water ASPs (set 3) had minimal effect on the correlation observed with the uncorrected values (Figure 3.9B). The range of the calculated interaction energies increased from 73 to 81 kcal mol^{-1} and the relative rank of one of the complexes shifted relative to the others. ASP sets 5 and 8, in contrast, had very deleterious effects on the calculated interaction energies (Figures 3.9C and 3.9D). The correlation between the experimental free energies and the calculated interaction energies became weaker in both cases, especially in the case of the vapor-water-derived ASP set 8. One of the calculated interaction energies for ASP set 5 (Figure 3.9C) and all of those for ASP set 8 (Figure 3.9D) became positive. The relative ranking of the complexes was most significantly disturbed by ASP set 8 (Figures 3.9A - 3.9D).

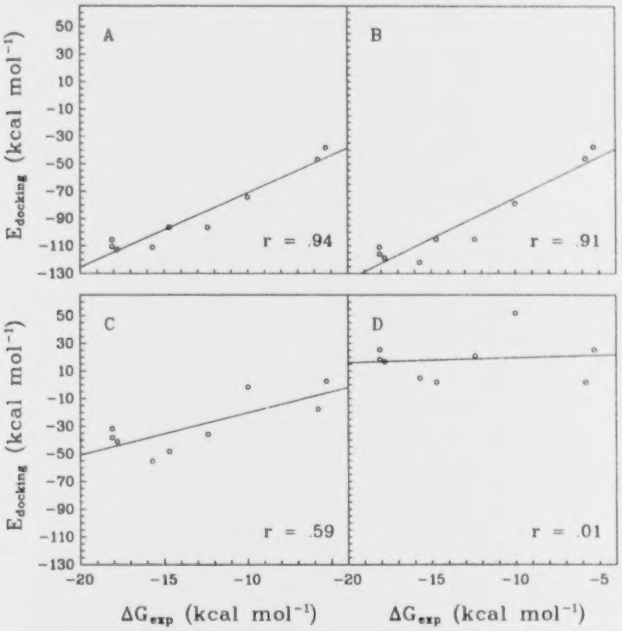


Figure 3.9: Correlation of calculated interaction energies and experimental free energies for a series of proteinase-inhibitor complexes. The correlation coefficients for each plot are shown in the bottom right-hand corner of each plot. **A-D:** as in Figure 3.3. Complexes used: 2tpi, ΔG_{exp} = -5.8 and -18.1 kcal mol^{-1}, the latter is the upper point with ΔG_{exp} = -18.1 kcal mol^{-1} in all 4 plots; 3sgb, ΔG_{exp} = -14.7 kcal mol^{-1}; 2kai, ΔG_{exp} = -12.4 kcal mol^{-1}; 3cpa, ΔG_{exp} = -5.3 kcal mol^{-1}; 1cho, ΔG_{exp} = -15.7 kcal mol^{-1}; 2ptc, ΔG_{exp} = -18.1 kcal mol^{-1}; 1tpa, ΔG_{exp} = -17.8 kcal mol^{-1}; 4cpa, ΔG_{exp} = -10.0 kcal mol^{-1}.

3.3.6 Evaluation of ASPs: summary

It is clear from the results presented here that the ASP set most consistent with, or complementary to, the standard Lennard-Jones plus Coulomb potential function used in BOXSEARCH is that derived from octanol-water transfer energies (ASP set 3). The ASPs derived from other solvent sytems disturbed the relative ranking of dockings in many of our test systems. These results are consistent with those reported recently by other workers (Schiffer et al., 1993). Also, the "entropy-corrected hydrophobic energies" derived by Abagyan & Totrov (1994), albeit in a somewhat arbitrary fashion [optimization to follow (Abagyan & Totrov, 1994)], seem to be most similar in magnitude to the octanol-water-based desolvation corrections reported here. Our results, and those of Schiffer *et al.* (1993), are also consistent with the theoretical considerations of Ben-Naim (1994) regarding the importance of the protein backbone in the processes of solvent-solvent transfer and protein desolvation. Only the octanol-water transfer energies are derived from experiments with sidechains that are attached to a peptide-like backbone, and this ASP set is the best complement to a standard potential function. While Wolfenden's group has shown that the relative hydrophobicities of Trp and Phe are not affected by the presence or absence of a one-residue backbone (Radzicka & Wolfenden, 1988), this result has not been shown to be generally applicable to peptide and protein desolvation. Other work has established that the solvation energies of small fragments are conditional upon the molecule to which thay are attached (Ben-Naim, 1993).

The complementarity of the octanol-water ASPs to a standard potential function may indicate that the octanol environment resembles a protein-protein interface more closely than do cyclohexane or vapor. This seems reasonable given the presence of the alcohol hydroxyls and the significant amount of water present in wet octanol (Radzicka & Wolfenden, 1988). Thus, bulk octanol contains significant proportions of hydrophobic, hydrogen-bond-donating, and hydrogen-bond-accepting surface. Furthermore, these groups are mobile, so the surfaces can rearrange themselves to optimize interaction with the small amino acid analogs used in the transfer experiments. This is reminiscent of a protein surface that is complementary to a ligand, unlike either bulk cyclohexane or dilute vapor. The protein-protein complexes we have studied have interfaces that are 60-70% hydrophobic (C) and 5-10% polar (N/O). For octanol-water ASPs the relative weighting of the various parameters (Table 3.4) is such that the energetic reward for burying hydrophobic surface outweighs the penalty for burying the polar and charged surfaces. For correct dockings, which typically have larger interfaces and a higher proportion of hydrophobic surface [unpublished observations, and see also Chapter 2 and (Cummings et al., 1995)], this energetic reward is greater, and thus we are better able to distinguish correct from incorrect dockings. ASP sets derived from other solvent systems place much more emphasis on the polar and charged surfaces (Table 3.4). This probably reflects the absence of hydrogen bond donors and acceptors in the solvent systems from which these ASPs are derived. For all dockings the energetic penalty for burying polar and charged surfaces dominates the desolvation energy contribution. The corrected interaction energies obtained with these ASPs do not correlate as well with docking correctness as those obtained with octanol-water ASPs.

In their review of subunit-subunit interfaces, Janin et al. (1988) noted that Leu and Arg each contribute approximately 10% of the total surface area of the surveyed interactions. Although this observation may not apply to protein-protein interfaces in general, it suggests a useful comparison for the present discussion. Given the propensity of Leu and Arg to contribute to subunit-subunit interfaces, it seems likely that burial of these sidechains is, in general, energetically favorable. We estimate that the BOXSEARCH potential function will yield, at best, a binding energy of -12 kcal mol^{-1} for complete burial of a fully exposed Leu sidechain and, similarly, -15 kcal mol^{-1} for an Arg sidechain (Figure 3.3 shows that BOXSEARCH overestimates binding energies for protein-protein complexes by a factor of approximately six). After correction with any of the ASP sets, burial of Leu will still be energetically favorable, but the use of ASP set 8 will make this process approximately 8 kcal mol^{-1} less favorable than with ASP sets 3 or 5. On the other hand, ASP sets 3, 5, and 8 yield corrections of +2, +15, and +22 kcal mol^{-1}, respectively, for complete burial of a fully exposed Arg sidechain. Correction of the BOXSEARCH potential for desolvation will therefore make burial of Arg highly favorable with ASP set 3, neutral with ASP set 5, and highly unfavorable with ASP set 8. This example is consistent with our docking analysis results, and serves to underline the appropriateness of ASP set 3 for our purpose.

3.4 Conclusion

While much success has been achieved with ASP-based empirical solvation/desolvation corrections (citations in the Introduction), it is important to remember that the correction described by equation (3.2) is a crude model of some aspects of the solvation/desolvation processes involved in protein folding and intermolecular interactions. As many workers have noted, it is unlikely that any single organic solvent is a good model for the environment encountered by an atom buried in protein (Wolfenden et al., 1981; Eisenberg & McLachlan, 1986). Furthermore, the behavior of a small amino acid analog in these transfer systems may not be entirely representative of the influence that that same residue would exert on the behavior of an extended peptide, or folded protein, in a similar transfer system. Ben-Naim has discussed in some detail why the empirical relationship between the surface areas and transfer energies of small blocked amino acids cannot be extended in an exact way to proteins (Ben-Naim, 1993; Ben-Naim, 1994). For example, this simple additive approach takes no account of the effects of intramolecular interactions between proximal sidechains. Also, this method introduces a "double-counting" error (van der Waals interaction of the solute with the non-polar solvent in the transfer system, and the van der Waals energy calculated with our potential function) into our corrected potential function (Ben-Naim, 1993).

In some of the sytems we tested, the improvement in ranking obtained with application of even the most suitable ASP set was small. In these cases our standard Lennard-Jones plus Coulomb potential function ranks the low energy dockings quite well without a desolvation correction. From a pragmatic and empirical perspective we are interested in any method that improves our ability to rank dockings correctly. Although our results suggest that an octanol-water-based ASP set is helpful, so far we do not see a uniform improvement that is independent of the docking system being studied. While it is, in general, better to include a desolvation correction using ASP set 3 than to omit it, there is clearly a great deal of room for improvement to the theory and practice of such corrections.

3.5 References

Abagyan, R. & Totrov, M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, 235:983–1002.

Abraham, D. J. & Leo, A. J. (1987). Extension of the fragment method to calculate amino acid zwitterion and side chain partition coefficients. *Proteins: Structure, Function, and Genetics*, 2:130–152.

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Prot. Eng.*, 2:101–113.

Bacon, D. J. & Moult, J. (1992). Docking by least-squares fitting of molecular surface patterns. *J. Mol. Biol.*, 225:849–858.

Ben-Naim, A. (1993). Solvation thermodynamics of biopolymers. In Westhof, E., editor, *Water and biological macromolecules*, pages 430–459. CRC Press.

Ben-Naim, A. (1994). Solvation: from small to macro molecules. *Curr. Top. Struc. Biol.*, 4:264–268.

Ben-Naim, A. & Mazo, R. M. (1993). Size dependence of the solvation free energies of large solutes. *J. Phys. Chem.*, 97:10829–10834.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542.

Cherfils, J., Bizebard, T., Knossow, M., & Janin, J. (1994). Rigid-body docking with mutant constraints of influenza hemagglutinin with antibody HC19. *Proteins: Structure, Function, and Genetics*, 18:8–18.

Cherfils, J., Duquerroy, S., & Janin, J. (1991). Protein-protein docking analyzed by docking simulation. *Proteins: Structure, Function, and Genetics*, 11:271–280.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248:338–339.

Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature*, 256:705–708.

Clackson, T. & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, 267:383–386.

Cook, W. J., Jeffrey, L. C., Carson, M., Chen, Z., & Pickart, C. M. (1992). Structure of a diubiquitin conjugate and a model for interaction with ubiquitin conjugating enzyme (E2). *J. Biol. Chem.*, 267:16467–16471.

Cummings, M. D., Hart, T. N., & Read, R. J. (1995). Monte Carlo docking with ubiquitin. *Protein Sci.*, 4:885–899.

Delbaere, L. T. J., Hutcheon, W. L. B., James, M. N. G., & Thiessen, W. E. (1975). Tertiary structure differences between microbial serine proteases and pancreatic serine enzymes. *Nature*, 257:758–763.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29:7133–7155.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, 319:199–203.

Eisenberg, D., Wesson, M., & Yamashita, M. (1989). Interpretation of protein folding and binding with atomic solvation parameters. *Chemica Scripta*, 29A:217–221.

Evans, S. V. (1993). Setor: hardware lighted three-dimensional solid model representations of macromolecules. *J. Mol. Graphics*, 11:134–138.

Fan, W., Tsai, R.-S., El Tayar, N., Carrupt, P.-A., & Testa, B. (1994). Solute-water interactions in the organic phase of a biphasic system. 2. Effects of organic phase and temperature on the "water-dragging" effect. *J. Phys. Chem.*, 98:329–333.

Fauchère, J.-L. & Pliska, V. (1983). Hydrophobic parameters π of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.*, 18:369–375.

Guy, H. R. (1985). Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophys J.*, 47:61–70.

Hart, T. N. & Read, R. J. (1992). A multiple-start Monte Carlo docking method. *Proteins: Structure, Function, and Genetics*, 13:206–222.

Hart, T. N. & Read, R. J. (1994). Multiple-start Monte Carlo docking of flexible ligands. In Merz, K.M., Jr. & LeGrand, S., editors, *The Protein Folding Problem and Tertiary Structure Prediction*, pages 71–108. Birkhäuser, Boston.

Holtzer, A. (1994). Does Flory-Huggins theory help in interpreting solute partitioning experiments. *Biopolymers*, 34:315–320.

Horton, N. & Lewis, M. (1992). Calculation of the free energy of association for protein complexes. *Prot. Sci.*, 1:169–181.

Janin, J., Miller, S., & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.*, 204:155–164.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Prot. Chem.*, 14:1–63.

Korn, A. P. & Burnett, R. M. (1991). Distribution and complementarity of hydropathy in multisubunit proteins. *Proteins: Struct. Funct. Genet.*, 9:37–55.

Ooi, T., Oobatake, M., Némethy, G., & Scherage, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Nat. Acad. Sci., U.S.A.*, 84:3086–3090.

Radzicka, A. & Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, 27:1664–1670.

Radzicka, A., Young, G. B., & Wolfenden, R. (1993). Lack of water transport by amino acid side chains or peptides entering a nonpolar environment. *Biochemistry*, 32:6807–6809.

Read, R. J., Fujinaga, M., Sielecki, A. R., & James, M. N. G. (1983). Structure of the complex of *Streptomyces griseus* protease B and the third domain of the turkey ovomucoid inhibitor at 1.8 Å resolution. *Biochemistry*, 22:4420–4433.

Richmond, T. J. (1984). Solvent accessible surface area and excluded volume in proteins. *J. Mol. Biol.*, 178:63–89.

Rose, G. D. & Wolfenden, R. (1993). Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu. Rev. Biophys. Biomol. Struct.*, 22:381–415.

Sawyer, L., Sielecki, A. R., & James, M. N. G. unpublished coordinates.

Schiffer, C., Caldwell, J. W., Kollman, P. A., & Stroud, R. M. (1993). Protein structure prediction with a combined solvation free energy- molecular mechanics force field. *Molec. simulation*, 10:121–149.

Sharp, K. A., Nicholls, A., Friedman, R., & Honig, B. (1991). Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models. *Biochemistry*, 30:9686–9697.

Shoichet, B. K. & Kuntz, I. D. (1991). Protein docking and complementarity. *J. Mol. Biol.*, 221:327–346.

Shrake, A. & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. lysozyme and insulin. *J. Mol. Biol.*, 79:351–371.

Still, W. C., Tempczyk, A., Hawley, R. C., & Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129.

Stouten, P. F. W., Frommel, C., Nakamura, H., & Sander, C. (1993). An effective solvation term based on atomic occupancies for use in protein simulations. *Molec. simulation*, 10:97–120.

Totrov, M. & Abagyan, R. (1994). Detailed ab initio prediction of lysozyme-antibody complex with 1.6 Å accuracy. *Struc. Biol.*, 1:259–263.

Tsai, R.-S., W., F., El Tayar, N., Carrupt, P.-A., Testa, B., & Kier, L. B. (1993). Solute-water interactions in the organic phase of a biphasic system. 1. Structural influence of organic solutes on the "water-dragging" effect. *J. Am. Chem. Soc.*, 115:9632–9639.

Vakser, I. A. & Aflalo, C. (1994). Hydrophobic docking: A proposed enhancement to molecular recognition techniques. *Proteins: Struct. Funct. Genet.*, 20:320–329.

Vijay-Kumar, S., Bugg, C. E., & Cook, W. J. (1987). Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.*, 194:531–544.

von Freyberg, B., Richmond, T. J., & Braun, W. (1993). Surface area included in energy refinement of proteins. *J. Mol. Biol.*, 233:275–292.

Wesson, L. & Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.*, 1:227–235.

Williams, R. L., Vila, J., Perrot, G., & Scheraga, H. A. (1992). Empirical solvation models in the context of conformational searches: application to bovine pancreatic trypsin inhibitor. *Proteins: Structure, Function, and Genetics*, 14:110–119.

Wodak, S. J. & Janin, J. (1978). Computer analysis of protein-protein interactions. *J. Mol. Biol.*, 124:323–342.

Wolfenden, R., Andersson, L., Cullis, P. M., & Southgate, C. C. B. (1981). Affinities of amino acid side shains for solvent water. *Biochemistry*, 20:849–855.

Young, L., Jernigan, R. L., & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Prot. Sci.*, 3:717–729.

Chapter 4

Modeling the carbohydrate-binding specificity of pig edema toxin[1]

4.1 Introduction

The Shiga-like toxins (SLTs), also known as verotoxins (VTs), are a class of bacterial toxins expressed by certain *Escherichia coli* serotypes known to cause several human and animal diseases. Members of this group include verotoxins 1 (SLT-I, VT-1) and 2 (SLT-II, VT-2), pig edema toxin (SLT-IIe, VT-2e, VT-E), and Shiga-like toxin IIc [SLT-IIc, VT-2c; (Lingwood, 1993)]. They have a hexameric AB_5 subunit composition: A represents the catalytically active monomeric subunit that ultimately elicits the toxic effect, and B_5 denotes the homopentameric binding subunit. The binding subunits of these toxins facilitate entry into certain host cells, by virtue of their ability to bind to specific host cell glycolipids.

---

[1] Manuscript in preparation.

There is greater than 60% sequence identity between the binding subunits of all verotoxins (Lingwood, 1993). SLT-I, SLT-II, and SLT-IIc bind globotriaosylceramide (Gb3; αGal(1-4)βGal(1-4)βGlcCer, where Cer is ceramide) preferentially, and the prevalence of this glycolipid in kidney tissue contributes to their renal toxicity (Lingwood, 1993). The binding subunit of SLT-IIe, which shows 84% sequence identity with that of SLT-II, prefers to bind globotetraosylceramide (Gb4; βGal*N*Ac(1-3)αGal(1-4)βGal(1-4)βGlcCer), although it does bind Gb3 as well (Lingwood, 1993). Brunton and co-workers explored these differences in binding preference by constructing several mutants of the SLT-I and SLT-IIe binding subunits (Tyrrell et al., 1992). One SLT-IIe double mutant (GT3; Gln65→Glu, Lys67→Glu) was particularly interesting, in that its binding preference was switched from Gb4 to Gb3 (Tyrrell et al., 1992). Later it was shown that this change in binding preference gave rise to a corresponding change in the *in vivo* activity of the toxin (Boyd et al., 1993). GT3, with Gb3/Gb4 binding activity similar to that of SLT-I, had a similar pathology to that of SLT-I, and this was distinct from that of wild-type SLT-IIe (Boyd et al., 1993).

Structural studies of these lectins, both free and complexed with their carbohydrate receptors, will help in understanding the binding interaction at the atomic level. In turn, such knowledge may aid in the development of vaccines and/or chemotherapeutic agents that block toxin binding. Armstrong and co-workers have recently demonstrated the feasibility of such an approach to the diagnosis and/or treatment of enterohemorrhagic *E. coli* infections (Armstrong et al., 1991). Hol and colleagues are compiling a body of structural information useful as a basis for the development of a similar approach to the treatment of cholera [(Merritt et al., 1994a) and refs therein]. Previous reports from this laboratory have described the crystallographic structure of the binding subunit of SLT-I

(Stein et al., 1992), and its structural relationship to the binding subunit of the cholera toxin family (Sixma et al., 1993). Nyholm et al. (1995) modeled the binding of the Gb3 to one site[2] of SLT-I, and postulated a second site of interaction. More recently, workers in this laboratory have reported the structure of SLT-I with a Gb3 analog bound at three distinct sites (Ling et al., 1995), and the determination of the structure of GT3 complexed with the same Gb3 analog is also underway (Ling et al., unpublished results).

Since Gb4 differs from Gb3 only by the addition of a terminal *N*-acetylgalactosamine moiety, it is possible that Gb3 and Gb4 bind to SLT-IIe with similar binding modes. Related similarities in the binding of various carbohydrates have been observed in structural studies of the heat-labile enterotoxin of *E. coli* (Merritt et al., 1994b). If this is the case then the loss of Gb4 affinity observed for the SLT-IIe double mutant GT3 may be due to the gain or loss of one or a few specific interactions with the mutated residues (DeGrandis et al., 1989; Tyrrell et al., 1992), or proximal sidechains that interact with either or both of the mutated residues. Here we report the results of modeling studies that provide an explanation for the difference in carbohydrate-binding preferences of SLT-IIe and the double mutant GT3, as well as predicting the interaction of SLT-IIe with Gb4.

4.2 Materials and methods

4.2.1 Protein structure preparation

The structures of the binding pentamers of SLT-I and (the SLT-IIe double mutant) GT3 in complex with a Gb3 analog (the carbohydrate moiety of the analog is identical to that of Gb3) are currently being refined in this laboratory (Ling et al., unpublished results). The SLT-I complex has four pentamers in the asymmetric unit; the GT3 complex has one. We arbitrarily chose pentamer 1 (VBA1-VBE1) of the SLT-I structure for our studies. For both SLT-I and GT3 we used a trimer consisting of monomers A, D, and E of the binding pentamer for all of our comparisons and minimizations, to reduce the size of the system being studied. The binding sites of monomer E were studied, with the assumption that the binding sites of this monomer would, in the presence of the neighbouring subunits A and D, behave similarly to those of this monomer in the pentamer.

---

[2] The numbering of the binding sites in the present work is different from that of Nyholm et al. (1995).

Protein structures were initially prepared according to the general method employed in our docking studies (Hart & Read, 1992; Cummings et al., 1995b). Monomers B and C, water molecules, and bound carbohydrates were deleted. Hydrogens were added to the protein in the standard way with INSIGHTII (Biosym Technologies, San Diego, California) at neutral pH. Polar hydrogen positions were then optimized with NETWORK (Bass et al., 1992), which maximizes hydrogen-bonding networks. This was followed by a round each of steepest descents and conjugate gradient minimization (200 steps maximum in each round; heavy atoms fixed) with the CVFF potential function of DISCOVER.

To construct the initial wild-type version of pig edema toxin from the crystal structure of the double mutant GT3 we simply converted the two mutant residues to those of wild-type SLT-IIe (Glu65→Gln, Gln67→Lys) using the standard procedure in INSIGHTII (Biosym Technologies, San Diego, California). For the initial model we selected the rotamers (using the rotamer library of Ponder & Richards (1987), as implemented in INSIGHTII) that were most similar to the conformations observed for the two mutant residues in the crystal structure. These sidechains were adjusted during manual docking, and were also unconstrained during the subsequent energy minimizations of the complexes.

The three Gb3 binding sites observed in the complex with SLT-I were reproduced in the SLT-IIe model by superimposing SLT-I onto GT3 (see below). Coordinates were superimposed and compared using unpublished programs written by Trevor Hart. Figures 4.2 and 4.4-4.6 were prepared with the MOLSCRIPT program (Kraulis, 1991).

4.2.2 Calculation of carbohydrate conformations

Starting models of the carbohydrate portions of Gb3 and Gb4 were generated by a combination of simulation methods, similar to one of the procedures suggested by Tvaroška and Pérez (1986). Rigid grid searches were carried out with the hard-spheres program GEGOP (Stuike-Prill & Meyer, 1990) to find low energy glycoside conformations for each of the two relevant disaccharides, lactose (βGal(1-4)βGlc) and galabiose (αGal(1-4)βGal), of Gb3. In these searches ϕ and ψ were searched in ten degree steps, and the glycosidic bond angle and the C5-C6 torsion were optimized at each step. We then selected several low energy conformations for each of the two disaccharides from the ϕ/ψ plots of these grid searches. After specifying the starting values for ϕ and ψ for each of these low energy conformations, the glycosidic linkage, the C5-C6 torsion, and all hydroxyl groups were minimized with GEGOP. Following this minimization we constructed the possible Gb3 conformers from the final disaccharide conformations and repeated the final minimization (as above for the disaccharides) with GEGOP. The resultant Gb3 structures were then imported into the DISCOVER (Biosym Technologies, San Diego, California) version of AMBER (Weiner et al., 1984), with the additional carbohydrate parameters developed by Homans (Homans, 1990). Charges and atom types were assigned as described by Homans (Homans, 1990). All heavy atoms were fixed, and hydrogen positions were minimized by the steepest descents method for a maximum of 200 steps. Following this, all atoms were optimized for a maximum of 500 steps of conjugate gradients minimization. For the AMBER minimizations a distance-dependent dielectric (ϵ=4r) was used, as described in the program documentation (Biosym Technologies, San Diego, California).

We generated three low energy conformers of the terminal disaccharide of Gb4 [βGal*N*Ac(1-3)αGal], as described above for the Gb3 disaccharides. Three Gb4 conformers were then generated by combining the Gb3 conformer of interest (see below) with each of the three low energy conformations of βGal*N*Ac(1-3)αGal. These three Gb4 structures were then minimized first with GEGOP, then with DISCOVER/AMBER, as described above for the Gb3 conformers.

4.2.3 Carbohydrate-protein complexes

The starting point for modeling the interaction of the Gb3 and Gb4 carbohydrates with wild-type SLT-IIe was the three distinct Gb3 binding sites modeled for monomer E of the binding subunit of SLT-IIe by superimposing the SLT-I complex onto SLT-IIe. Gb4 was initially modeled in these sites by superimposing the galabiose heavy atoms of Gb4 onto those of Gb3. Selected sidechains (see below), the exocyclic moieties of the *N*-acetylgalactosamine ring, and the glycosidic linkage of the βGal*N*Ac(1-3)αGal disaccharide of Gb4 were manually adjusted to optimize the protein-carbohydrate interaction and alleviate any bad contacts. These manually built complexes were then minimized with the DISCOVER version of AMBER, as described above for the carbohydrate structures. In some cases (see below) the cycle of manual model optimization followed by constrained energy minimization was repeated once or twice.

4.3 Results and discussion

4.3.1 Calculated Gb3 carbohydrate conformations

The ϕ/ψ plots of lactose and galabiose generated by the rigid grid searches with GEGOP are shown in Figure 4.1. For these simple disaccharides we expect this relatively crude hard-spheres method to give a reasonable estimate of the low energy conformers, although the energy barriers separating conformers will be over-estimated (Poppe et al., 1990b; French & Brady, 1990; Tvaroška & Pérez, 1986). We chose several conformers of each of the two disaccharides (Figure 4.1 legend) for further minimization with GEGOP (see Materials and methods). This more rigorous flexible

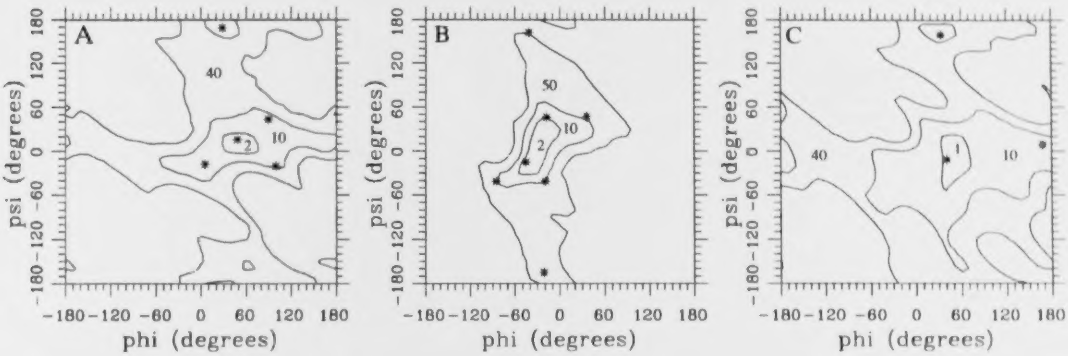


Figure 4.1: Conformational analysis of the glycosidic linkages of the constituent disaccharides of Gb3 and Gb4. Plots were obtained from the results of the original rigid grid searches with GEGOP, as described in the text. The contours contain the conformational space occupied by conformers that fall below the energy values (in kcal mol^{-1}) shown within the contours. Asterisks denote the conformers chosen for further refinement (described below). **A:** lactose; ϕ/ψ pairs chosen for further refinement with GEGOP were 50/10 (minimum energy), 100/-20, 90/40, 0/-20, and 30/170; the two final ϕ/ψ pairs used in Gb3 construction were 52/4 and 30/168. **B:** galabiose; ϕ/ψ pairs chosen for further refinement with GEGOP were -40/-20 (minimum energy), -40/160, -20/-160, -80/-40, -20/-40, -20/50, 30/50; the two final ϕ/ψ used in Gb3 construction were -40/-16 and -23/-157. **C:** βGal*N*Ac(1-3)αGal; ϕ/ψ pairs chosen for further refinement with GEGOP were 50/-10 (minimum energy), 170/10, 30/160; the three final ϕ/ψ pairs used in Gb4 construction were 54/-6, 170/-1, and 36/159.

stage of refinement[a]	energy	galabiose		lactose	
		ϕ^b	ψ^b	ϕ	ψ
	(kcal mol^{-1})	(degrees)			
conformer 1					
GEGOP	-4.7	-39.6	-16.3	52.9	3.6
DISCOVER	-8.9	-46.6	-9.7	49.1	1.5
conformer 2					
GEGOP	1.6	-39.7	-16.1	30.3	168.4
DISCOVER	-7.3	-46.6	-9.7	52.2	-164.2
conformer 3					
GEGOP	11.6	-23.2	-157.1	56.6	1.1
DISCOVER	-5.0	-20.3	-161.5	49.4	1.6
conformer 4					
GEGOP	18.3	-23.1	-156.9	31.9	168.1
DISCOVER	-3.5	-19.9	-163.8	52.0	-165.5

Table 4.1: Energies and glycoside conformations of low energy Gb3 conformers. [a]Energies and dihedral values refer to the final two stages of refinement of the modelled Gb3 conformations (see text). [b]ϕ: H1-C1-O1-Cx; ψ: C1-O1-Cx-Hx.

minimization led to quite drastic conformational adjustments in some cases, such that we were left with just two distinct ϕ/ψ conformers for each of the two disaccharides (Figure 4.1, Table 4.1). The four possible Gb3 conformers constructed by combining these disaccharide conformations were subjected to another round of flexible minimization with GEGOP, followed by minimization with DISCOVER/AMBER as described above. Neither of these final rounds of minimization with the Gb3 trisaccharides resulted in major conformational changes for any of the four conformers. Table 4.1 shows the ϕ and ψ angles and energy values for the Gb3 conformers at the two latter stages of refinement. We note that conformer 1, the lowest energy conformation, is similar to the carbohydrate moiety of the low energy Gb3 conformer modeled by Nyholm et al. (1995) in their docking study of Gb3 and SLT-I, and also resembles the lowest energy conformer described by Poppe et al. (1990b) in their combined NMR and computational analysis of Gb3 conformation.

4.3.2 Comparison of calculated and bound Gb3 conformations

There are three distinct Gb3 binding sites observed per monomer for the binding subunit of SLT-I, and one or two sites[3] per monomer for the binding subunit of GT3 (Figure 4.2; see below; Ling et al., unpublished coordinates). The ranges and average ϕ and ψ values of the Gb3 conformers bound at each of these four sites are shown in Table 4.2. For the SLT-I structure, the glycoside conformation of the lactose moieties show wide variation over the twenty copies of each of the three unique binding sites (Table 4.2). This is a reflection of the inherent flexibility of this glycosidic linkage [Figure 4.1; (Poppe et al., 1990b) and refs therein], lack of contact between the protein and the glucose residue, and, in the case of a few of the sites, crystal contacts. The average values of these parameters are, however, similar for the three SLT-I sites, as well as for the single GT3 site (Table 4.2; see Footnote 1). The galabiose moieties are much more restricted, reflecting the greater contact that this disaccharide makes with the protein. The conformation of the Gb3 carbohydrate bound at site I of SLT-I is most similar to that bound to GT3 (Table 4.2). Of the four low energy Gb3 conformers that we modeled, the one most similar to the bound conformers is

---

[3] At the time that this work was carried out, sugar binding had only been observed at site I of GT3. More recently, partial occupancy of site III of GT3 has been observed (Ling et al., unpublished coordinates).

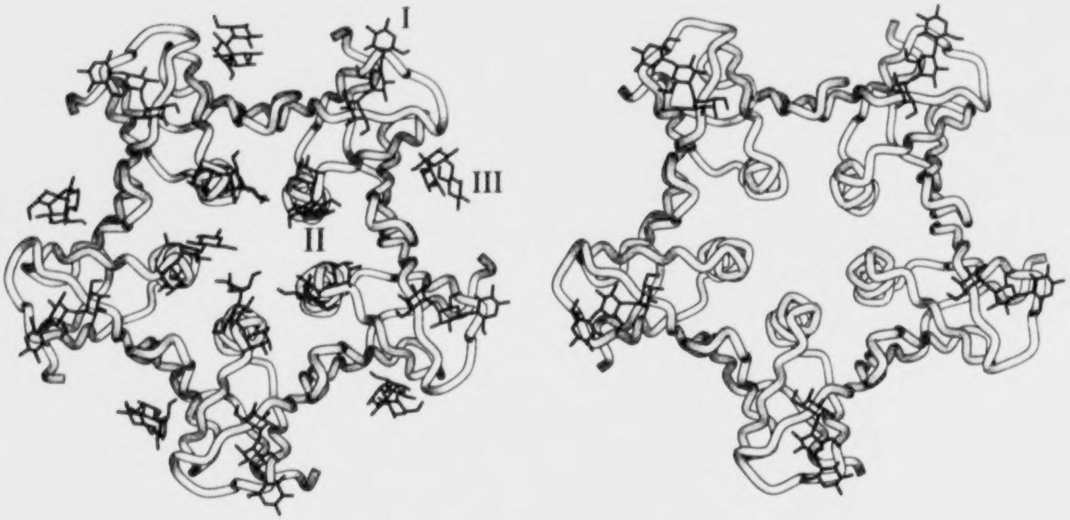


Figure 4.2: Crystallographic Gb3 binding sites of SLT-I and GT3. The three unique Gb3 binding sites of SLT-I (left) are shown, as well as the numbering used throughout the text. Gb3 binding site I of GT3 (right; see Footnotes 2 and 3) is also shown. The figure clearly shows the relatedness of the single Gb3 site observed initially for GT3 (see Footnote 1) and SLT-I site I.

conformer 1, the lowest energy conformer.

4.3.3 Gb3 binding sites on SLT-I and GT3

Although the proteins were crystallized under similar conditions, SLT-I shows three distinct Gb3 binding sites (Figure 4.2), whereas only one was observed initially for GT3 [Figure 4.2; see above and Footnote 2; to be discussed in detail elsewhere (Ling et al., manuscript in preparation)]. The biological significance of the different Gb3 binding sites observed for these proteins is not clear. There is a distinct possibility, previously noted by Tyrrell et al. (1992), that more than one unique

conformer	galabiose		lactose	
	ϕ[b]	ψ[b]	ϕ	ψ
	(degrees)			
GT3	-42.8 (-46.6, -40.4)	-17.1 (-17.9, -15.9)	38.3 (36.9, 39.4)	-23.4 (-25.7, -19.0)
SLT-I site I	-50.3 (-58.2, -37.3)	-11.1 (-20.5, 0.8)	45.2 (30.1, 70.3)	1.5 (-21.2, 36.2)
SLT-I site II	-43.4 (-48.9, -36.7)	-2.6 (-17.4, 11.6)	32.2 (-92.8, 78.0)	-8.8 (-60.9, 63.1)
SLT-I site III	-46.7 (-67.9, -13.4[c])	-13.6 (-34.4, 1.4)	57.6 (15.7, 139.8)	-8.5 (-87.6, 68.4)

Table 4.2: Glycoside conformations of Gb3 carbohydrates bound to GT3 and SLT-I. [a]Hydrogens were added in INSIGHTII in the standard way at neutral pH. [b]For SLT-I, values shown are the average of the twenty copies found in the asymmetric unit; for GT3, values shown are the average of the five copies found in the asymmetric unit; values in parentheses define the range of values observed for that parameter; ϕ and ψ defined in Table 1. [c]Only one value was below -36.5; omission of this point (-13.4) gave a mean of -48.4.

site may be biologically relevant. The calorimetry results of Toone and co-workers (St. Hilaire et al., 1994) indicate one unique binding site, for the Gb3 carbohydrate, per monomer of the SLT-I binding pentamer. While this result is quite convincing, it does not exclude the possibility of more than one Gb3/Gb4 binding site per monomer of the SLT-I and/or the SLT-IIe binding pentamer. Indeed, it may be that the Gb3 and Gb4 binding sites on these binding pentamers are distinct. The single Gb3 binding site observed initially for GT3 (see Footnotes 2 and 3) is common to both structures (site I in Figure 4.2), and the conformation of the Gb3 carbohydrate bound at this site is similar for both complexes (discussed above). However, it is not clear how the two mutations of GT3 would affect the binding of Gb3 and/or Gb4 at this site (see below). This leads us to consider that one of the other two Gb3 binding sites on SLT-I may be relevant to Gb4 binding by SLT-IIe [further refinement of the GT3 complex has revealed partial occupancy of site III of GT3 (Ling et al., unpublished coordinates); see Footnote 2].

Several residues of the binding subunits of SLT-I and SLT-IIe have been studied by site-directed mutagenesis, and some of these residues have been inferred to be involved in receptor binding (Jackson et al., 1990; Tyrrell et al., 1992; Clark et al., 1995). If slight alteration of a sidechain (*eg.* Val→Leu, Glu→Gln etc.) has a marked effect on ligand binding, and is shown not to affect significantly other important properties of the protein, then the simplest interpretation is that the sidechain in question interacts with the ligand. Although such reasoning is commonly invoked, in the absence of direct experimental verification of these interactions such inferences are always somewhat tenuous. This point is of some concern to us, since the distance constraints offered by specific intermolecular interactions are potentially of great utility in computer-based docking simulations. We have compiled some of the mutation data available for the binding subunits of SLT-I and SLT-IIe, in an effort to understand the relevance of the three Gb3 binding sites observed for SLT-I, and to help us understand the change in binding preference observed for the double mutant GT3.

We focused on the following residues of the binding subunit of SLT-IIe: Trp30, the Glu16-Asp17-Asn18 sequence, Gln65, and Lys67. The reader should note that we use the SLT-I numbering throughout the present report (shown in Figure 4.3). The Phe30→Ala substitution for the SLT-I binding subunit caused a marked reduction in Gb3 affinity, and protein structure was not affected by this substitution (Clark et al., 1995). Given the drastic nature of the Phe→Ala substitution, and the similar positions of the two residues, it is reasonable to expect a similar result for the Trp30→Ala substitution in SLT-IIe (Figure 4.3). The effects of mutation of the Asp16-Asp17-Asp18 sequence on Gb3 and Gb4 binding by SLT-I have been studied by

```
                10        20        30        40        50        60
                .         .         .         .         .         .
SLT-I: TPDCVTGKVEYTKYNDDDTFTVKVGDKELFTNRWNLQSLLLSAQITGMTVTIKTNACHNGGGFSEVIFR
         ||  || |  ||| | ||||||   |  ||||||| || ||| |||||||  | |  | || || | 
  GT3:  ADCAKGKIEFSKYNEDNTFTVKVSGREYWTNRWNLQPLLQSAQLTGMTVTIISNTCSSGSGFAEVQFN
                .         .         .         .         .         .    * *
                10        20        30        40        50        60   Q K
```

Figure 4.3: Alignment used for superposition of SLT-I onto GT3. Stars denote the two mutated residues of GT3, that we replaced in our model of SLT-IIe. The Cα's of the identical residues of SLT-I, denoted by vertical lines between the two sequences, were superimposed onto the corresponding atoms of GT3/SLT-IIe.

Tyrrell et al. (1992) and Jackson et al. (1990). Concomitant replacement of Asp16 and Asp17 by His abolished binding of a Gb3 analog, but the more conservative substitution Asp→Asn, singly (16, 17, and 18) or in pairs (16 and 17), had little or no effect (Jackson et al., 1990; possible explanations for these effects are discussed below). The Asp18→Asn mutation yields a SLT-I mutant which binds both Gb3 and Gb4 [instead of just Gb3; (Tyrrell et al., 1992)]. Tyrrell et al. (1992) noted that the reciprocal mutation of SLT-IIe (Asn18→Asp) did not affect Gb4 binding. This region of these proteins may be involved in receptor binding, and the differences between the SLT-I and SLT-IIe (Glu16-Asp17-Asn18) sequences may contribute to the differences in Gb3/Gb4 binding preferences of the two toxins. Residues Gln65 and Lys67 of SLT-IIe have been shown to be important for Gb4 binding (Tyrrell et al., 1992). The SLT-IIe double mutant GT3 (Gln65→Glu, Lys67→Gln) had a markedly reduced affinity for Gb4, but affinity for Gb3 was moderately increased (Tyrrell et al., 1992). A reasonable conclusion is that the Gal*N*Ac moiety of Gb4 makes specific favorable contacts with Gln65 and Lys67, and that these contacts are not possible for the double mutant GT3 (Tyrrell et al., 1992).

To generate the three SLT-IIe Gb3 binding sites from those observed for SLT-I (Figure 4.2), we superimposed the Cαs of selected residues of monomer E of SLT-I onto the corresponding atoms of SLT-IIe (43 residue pairs, see Figure 4.3). The final RMS difference for the superposed atoms was 0.39 Å.

Figure 4.4 shows the central monomer (monomer E) of the SLT-IIe binding subunit trimer used in our modeling experiments, with the residues discussed above highlighted. Also shown are the three Gb3 "binding sites" of monomer E of SLT-IIe, modeled as described above. Visual inspection of this model shows clearly that the implicated residues (particularly the two residues mutated in GT3) cluster most tightly around site III. Assuming a similar binding mode for the analogous portions of Gb3 and Gb4, both ligands would have the greatest interaction with these residues when bound in site III (Figure 4.4). Since a direct interaction with the mutated residues offers the simplest explanation of the change in affinity, we decided to explore Gb4 binding in site III of our modeled wild-type protein.

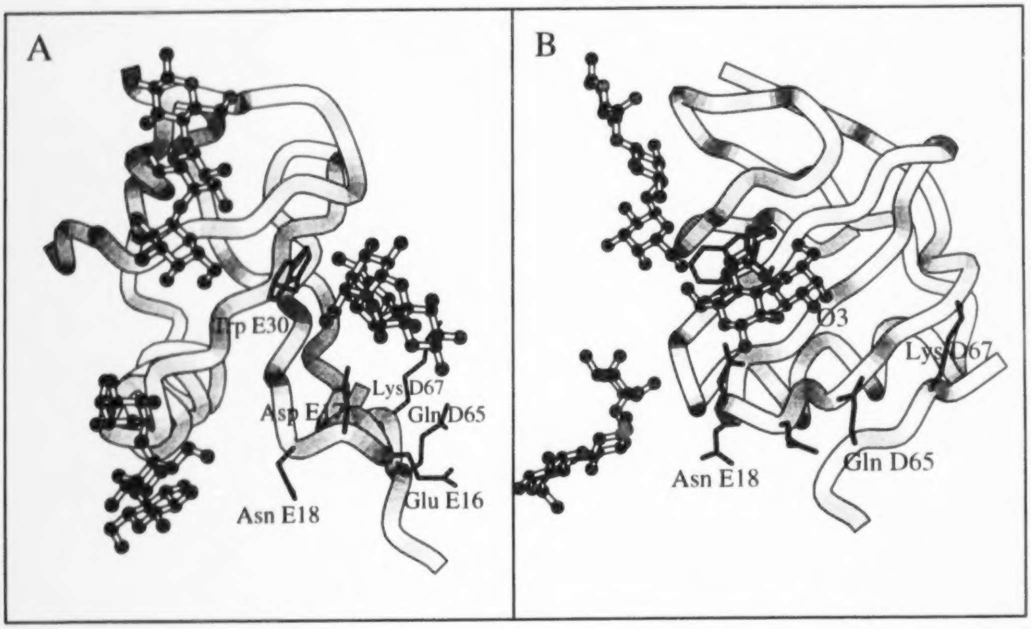


Figure 4.4: Initial model of SLT-IIe with three unique Gb3 binding sites. The backbone of monomer E, and residues D63-D69, is included, with the sidechains of several key residues (see text) shown. **B** is a 60° rotation of **A**. In **B**, O3 of the terminal Gal residue of Gb3, the Gal*N*Ac linkage point for Gb4, is labelled for the Gb3 molecule bound at site III.

4.3.4 Gb4 conformations

We assumed that the Gb3 moiety of Gb4 would bind at site III (discussed above) of our modeled wild-type pig edema toxin (SLT-IIe), in a mode similar to that of Gb3 bound at the analogous site in the SLT-I complex. Thus, to generate reasonable conformers of Gb4 for modeling the Gb4/pig edema toxin site III complex, we had only to concern ourselves with the terminal βGal*N*Ac(1-3)αGal glycosidic linkage.

Figure 4.1 shows the original ϕ/ψ plot for this disaccharide generated with the GEGOP rigid grid search (as described for lactose and galabiose), and the three conformers chosen for further minimization with GEGOP. Table 4.3 lists the ϕ and ψ angles and energies of the three Gb4 carbohydrate conformers constructed by combining Gb3 conformer 1 (Table 4.1) and the three refined βGal*N*Ac(1-3)αGal

		GalNAcβ1-3Gal		galabiose		lactose	
stage of refinement[a]	energy (kcal mol^{-1})	ϕ[b]	ψ[b]	ϕ	ψ	ϕ	ψ
				(degrees)			
conformer 1							
GEGOP	-6.8	51.8	2.7	-39.5	-16.2	53.5	3.4
DISCOVER	-25.3	53.3	-8.0	-43.2	-2.2	49.8	2.2
conformer 2							
GEGOP	-6.9	168.3	0.0	-39.8	-17.7	53.1	-1.0
DISCOVER	-27.6	175.6	-3.6	-46.6	-9.8	49.1	0.5
conformer 3							
GEGOP	-0.6	36.8	158.3	-39.6	-16.2	53.2	3.0
DISCOVER	-23.2	36.3	159.0	-47.0	-11.6	49.2	1.1

Table 4.3: Energies and glycoside conformations of low energy Gb4 conformers. [a]Energies and dihedral values refer to the final two stages of refinement of the modelled Gb4 conformations (see text). [b] See Table 4.1

conformers (Figure 4.1), following the final two rounds of energy minimization with GEGOP and DISCOVER/AMBER (as described above for the Gb3 conformers).

The final conformation of the Gb3 moiety of the three calculated Gb4 conformers (essentially Gb3 conformer 1) is in good agreement with the results of Poppe et al. (1990a,b). However, none of the three βGal*N*Ac(1-3)αGal conformations can be considered identical to those of the low energy Gb4 conformers described by Poppe et al. (1990a). In this respect conformer 1 is most similar to the most popular low energy Gb4 conformer(s) of Poppe et al. (1990a); ϕ and ψ differing by approximately 20° and 40°, respectively.

4.3.5 Gb4 binding at site III of SLT-IIe

Our modeling of the Gb4/SLT-IIe complex commenced with the superposition of the galabiose moiety of our three calculated Gb4 conformers onto the corresponding fragment of Gb3 in site III of SLT-IIe. Figure 4.5[4] shows that the Gal*N*Acβ1-3Gal moiety of conformer 2 projects away from the protein[2]. Without major conformational changes of both protein and carbohydrate, close contact between the Gal*N*Ac ring of Gb4 and either of residues GlnD65 or LysD67 seems unlikely for this Gb4 conformation. Conformers 1 and 3 bound at this site seem more promising, although both clash with the sidechain of LysE13[5].

We subjected each of the three complexes to energy minimization with DISCOVER/AMBER, using an initial 100 steps of highly constrained steepest

---

[4] The reader should note that the *post*-minimization complexes are shown in Figure 4.5. However, minimization did not significantly affect the major *differences* between the three complexes. Projection of the Gal*N*Ac residue away from the protein in this complex is a result of the conformation of the βGal*N*Ac(1-3)αGal glycosidic linkage in this Gb4 conformer.

[5] As noted above, the *post*-minimization complexes are shown in Figure 4.5, and the clashes with LysE13 have largely been eliminated. It is clear, however, that LysE13 is near to the Gal*N*Ac ring in the complexes of conformers 1 and 3, but not in that of conformer 2.

initial steepest descents		final conjugate gradients	
residues[a]	constraint[b]	residues	constraint
Gb4:1	none	Gb4:1	none
–	–	Gb4:2-4	none
SLT-IIe:D53	bb	SLT-IIe:D53	bb
SLT-IIe:D65	bb	SLT-IIe:D65	bb
SLT-IIe:D67	bb	SLT-IIe:D67	bb
SLT-IIe:E13	bb	SLT-IIe:E11-E24	bb
SLT-IIe:E15	bb	–	–
–	–	SLT-IIe:E27-E31	bb
–	–	SLT-IIe:E58	bb
–	–	SLT-IIe:E59-E61	none

Table 4.4: Constraints for minimization of Gb4/SLT-IIe site III complexes. [a]All atoms were constrained to their initial positions, except those listed here. Initial Gb4/SLT-IIe complexes were minimized first with a more highly constrained round of steepest descents minimization (intitial steepest descents) with DISCOVER/AMBER, followed by a longer and less constrained round of conjugate gradients minimization (final conjugate gradients) with the same potential function (see text). [a]Gb4 is numbered progressively, with the Gal*N*Ac ring being residue 1. During steepest descents minimization the βGal*N*Ac(1-3)αGal glycosidic linkage was unconstrained. Each monomer of the SLT-IIe binding pentamer consists of 68 residues. The trimers used in our experiments were numbered A2-A69, D2-D69, and E2-E69, according to SLT-I numbering (see Figure 4.3). [b]None: unconstrained; bb: Cα, N, C, O atoms constrained.

descents minimization, followed by a more relaxed cycle of 300 steps of conjugate gradients minimization. Table 4.4 lists the constraints used in each of the two minimization protocols. Essentially, we allowed flexibility for sidechains and backbone regions in close contact with Gb3/Gb4 "bound" at this site. This included several residues previously implicated in receptor binding (discussed above). Our goal with the first round of minimization was to fix the Gb3 moiety of the Gb4s, as well as most of the protein, and allow the terminal Gal*N*Ac ring and selected parts of the protein to adjust to accomodate this docking. The less constrained second round of minimization should then allow further fine adjustment of slightly more of the

	Gb4 residue				
SLT-IIe[a] residue	GalNAc:1	Gal:2	Gal:3	Glc:4	total
Gb4 conformer 1					
Ile:D53	-0.5				-0.6
Gln:D65	-3.4				-3.6
Lys:D67	-1.6				-1.3
Lys:E13	-8.0				-8.1
Asn:E15	-1.5	-2.0			-3.8
Glu:E16	-1.0				-1.2
Asp:E17			-2.8	-0.8	-3.1
Thr:E19		-0.6	-1.2		-1.8
Thr:E21		-2.7			-3.2
Glu:E28		-3.4			-3.6
Tyr:E29		-0.5			-0.5
Trp:E30		-2.4	-4.8		-7.4
Gly:E60			-1.3		-1.8
total	-17.8	-13.5	-11.5	-1.0	-43.9
Gb4 conformer 2					
Lys:E13	-2.7				-2.7
Asn:E15		-2.3			-2.9
Asp:E17			-2.7	-0.7	-2.6
Thr:E19			-1.1		-1.6
Thr:E21		-2.1			-2.6
Glu:E28	-0.5	-4.9			-6.0
Tyr:E29		-1.1			-1.2
Trp:E30		-2.4	-4.9		-7.6
Gly:E60			-1.3		-2.1
total	-5.1	-15.3	-11.7	-1.1	-33.1
Gb4 conformer 3					
Gln:D65	-0.8				-1.0
Lys:E13		-1.1			-0.8
Asn:E15		-1.0	-1.0		-2.1
Asp:E17			-1.9	-0.8	-2.2
Thr:E19			-0.9		-1.2
Thr:E21		-2.2			-2.5
Glu:E28	-4.6	-3.5			-8.6
Tyr:E29		-1.6			-1.7
Trp:E30		-1.7	-4.7		-6.7
Gly:E60		-1.0	-1.4		-2.6
total	-8.4	-13.4	-11.3	-1.2	-34.2

Table 4.5: Important residue-residue interactions for the three initial complexes of Gb4 bound at SLT-IIe site III. Initial complexes following the first cycle of steepest descents/conjugate gradients minimization using the DISCOVER implementation of the AMBER potential function (see text). All residue-residue interactions with a total energy of -0.5 kcal mol^{-1} or less are listed, in units of kcal mol^{-1}. However, totals are the net total for that residue. [a]All of our simulations involved a trimer of subunits A, D, and E of the pentameric binding subunit, and we focused on the Gb3/Gb4 binding sites of monomer E. The three subunits of the trimer were numbered A2-A69, D2-D69, and E2-E69, according to SLT-I numbering (see Figure 4.3).

protein, as well as the complete Gb4 molecule. Figure 4.5 shows site III of each of the minimized complexes, and Table 4.5 lists the energies of the important favorable residue-residue interactions for each minimized complex.

Given our assumption (first discussed in general terms by DeGrandis et al., 1989, and later in more specific terms by Tyrrell et al., 1992) that the difference in binding preference between SLT-IIe and GT3 is due to interactions between the Gal*N*Ac moiety of Gb4 and GlnD65 and LysD67 (for site III of monomer E), the complex of SLT-IIe with Gb4 conformer 2 seems unlikely. This is somewhat surprising, since our calculations describe this as the most stable Gb4 conformation, at least of the three conformers we considered (Table 4.3; this differs slightly from the result of Poppe et al. (1990a), as discussed in the preceding section). In this complex Gal*N*Ac projects away from the protein surface (Figure 4.5), and most of this residue does not contact the protein. Only 242 Å^2 of non-polar surface (all carbons) is buried in this complex, compared to 291 Å^2 and 347 Å^2 for the conformer 1 and 3 complexes, respectively. The complex with Gb4 conformer 2 also involves the fewest (6, *versus* 9 and 7 for conformers 1 and 3, respectively) intermolecular hydrogen bonds of the three complexes. Table 4.5 shows that Gal*N*Ac is not a major contributor of binding energy for this complex. Of particular relevance is the observation that contacts between Gal*N*Ac and both GlnD65 or LysD67 are not important for this complex (Figure 4.5, Table 4.5). Thus, we see that conformer 2 does not provide a satisfactory model of Gb4 bound to (site III of) SLT-IIe.

The minimized complexes of Gb4 conformers 1 and 3 with SLT-IIe inspire more confidence. The complex of conformer 1 is favored over those of 2 and 3 by approximately 10 kcal mol^{-1} (Table 4.5). In both of these models (complexes with conformers 1 and 3) the Gal*N*Ac moiety projects toward the protein, and makes extensive contact with it (Figure 4.5). For conformer 3, contact between Gal*N*Ac and GlnE65 makes a relatively small favorable contribution to the stability of the

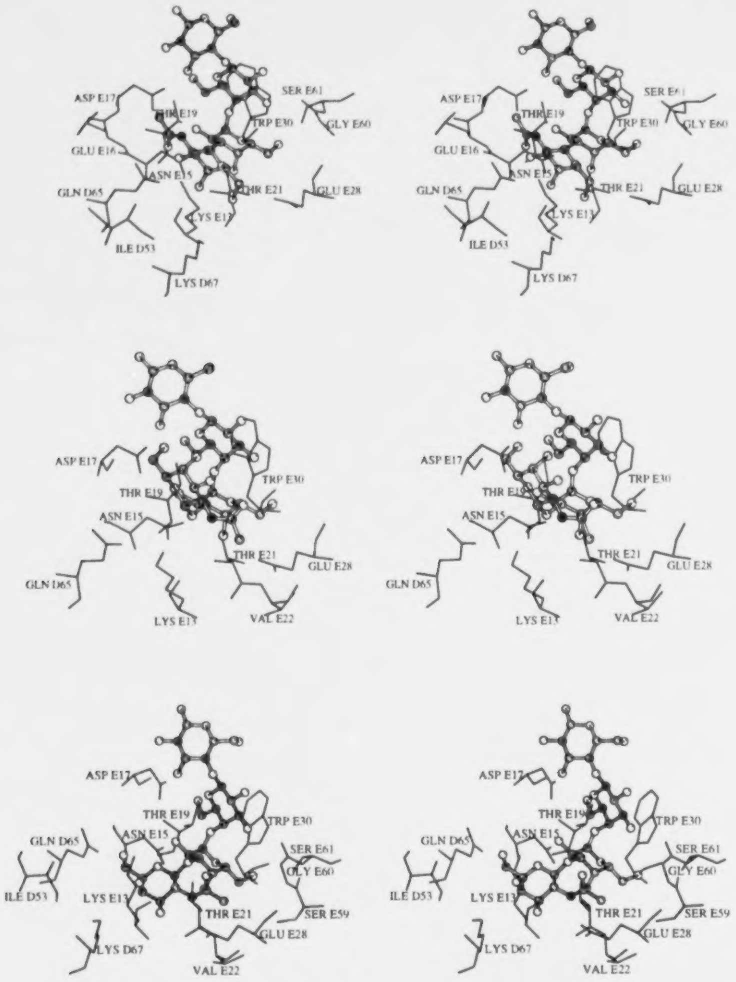

ASP E17
SER E61
THR E19
TRP E30
GLY E60
GLU E16
ASN E15
THR E21
GLU E28
GLN D65
LYS E13
ILE D53
LYS D67
ASP E17
SER E61
THR E19
TRP E30
GLY E60
GLU E16
ASN E15
THR E21
GLU E28
GLN D65
LYS E13
ILE D53
LYS D67
ASP E17
THR E19
TRP E30
ASN E15
THR E21
GLU E28
GLN D65
LYS E13
VAL E22
ASP E17
THR E19
TRP E30
ASN E15
THR E21
GLU E28
GLN D65
LYS E13
VAL E22
ASP E17
THR E19
TRP E30
GLN D65
ASN E15
SER E61
GLY E60
ILE D53
LYS E13
THR E21
SER E59
GLU E28
LYS D67
VAL E22
ASP E17
THR E19
TRP E30
GLN D65
ASN E15
SER E61
GLY E60
ILE D53
LYS E13
THR E21
SER E59
GLU E28
LYS D67
VAL E22

Figure 4.5: The minimized complexes of each of the three Gb4 conformers at site III of SLT-IIe. Most of the residues with one or more atoms within 5Å of a Gb4 atom are shown. Gb4 conformers I, II, and III are shown in the top, middle, and bottom figures, respectively.

complex (Table 4.5), and LysD67 does not contact Gb4 at all (Figure 4.5, Table 4.5). Compared to conformer 2, Gal*N*Ac makes a larger contribution towards the stability of this complex (Table 4.5). The complex with conformer 1, however, is by far the most attractive of the three models. The contribution of Gal*N*Ac towards the stability of this complex is greater than three times that for conformer 2 (Table 4.5), and more than twice that for conformer 3 (Table 4.5). The energetic consequence of the contacts between Gb4 and both GlnD65 and LysD67 (Table 4.5) also contributes to the attractiveness of this model.

All of our simulations were performed *in vacuo*, without any explicit consideration of solvent effects. A simple way of partially compensating for this deficiency is to apply a desolvation correction based on the area of non-polar surface buried upon complex formation [(Cummings et al., 1995a) and refs therein; (Eisenberg et al., 1989; Eisenberg & McLachlan, 1986; Chothia, 1974)]. If we ascribe an energetic reward of 20 cal $Å^{-2}$ of non-polar surface (all carbons), a compromise between several similar reported values [(Cummings et al., 1995a) and refs therein; (Eisenberg et al., 1989; Eisenberg & McLachlan, 1986; Chothia, 1974)], the complexes of conformers 1, 2, and 3, would be corrected by -5.8 kcal mol^{-1}, -4.8 kcal mol^{-1}, and -6.9 kcal mol^{-1}, respectively.

Following inspection of the minimized complexes of conformers 1 and 3 (Figure 4.5) we concluded that further optimization of the interactions would require manual intervention. We restricted ourselves to manipulation of the glycosidic linkage of βGal$N$Ac(1-3)αGal, the exocyclic N-acetyl group of the Gal$N$Ac ring, and the sidechains of LysD67 and LysE13 of the protein. The dihedrals of the sidechains were adjusted in steps of 120°. Figure 4.1 indicates that the βGal$N$Ac(1-3)αGal glycosidic linkage has considerable torsional flexibility in the regions occupied by both conformers 1 and 3 (Table 4.3).

For both of these minimized complexes the biggest problem was the prohibitive interaction with LysE13[3] (Figure 4.5). Minor manual adjustment of ϕ and ψ of βGal$N$Ac(1-3)αGal and the LysE13 and LysD67 sidechains appear to turn this problem to advantage for the complex of conformer 1 (Figure 4.6, Table 4.6). Such manipulations proved fruitless for the complex involving Gb4 conformer 3 (results not shown). Minor adjustments of these and other torsions did not yield a complex

	Gb4 residue				
SLT-IIe residue	Gal*N*Ac:1	Gal:2	Gal:3	Glc:4	total
Ile:D53	-0.6				-0.6
Gln:D65	-3.4				-3.6
Lys:D67	-3.5				-3.2
Lys:E13	-8.0				-8.1
Asn:E15	-1.6	-2.0			-3.8
Glu:E16	-1.0				-1.2
Asp:E17			-2.8	-0.8	-3.2
Thr:E19		-0.6	-1.2		-1.8
Thr:E21		-2.7			-3.2
Glu:E28		-3.3			-3.5
Tyr:E29		-0.5			-0.5
Trp:E30		-2.4	-4.8		-7.4
Gly:E60			-1.3		-1.8
total	-20.0	-13.4	-11.5	-1.0	-45.9

Table 4.6: Important residue-residue interactions for the final model of the complex of Gb4 bound at site III of SLT-IIe. The final modelled complex following initial minimization, manual adjustment, further minimization, manual adjustment, and a final round of minimization. All residue-residue interactions with a total energy of -0.5 kcal mol^{-1} or less are listed, in units of kcal mol^{-1}. However, totals are the net total for that residue (see text, and also Table 4.5).

involving direct interactions between Gal*N*Ac and the sidechains of GlnD65 and LysD67.

Two rounds of both minor manual adjustment (see above) and 100 steps of conjugate gradients minimization, with constraints as in step 2 of our original two-step minimization (Table 4.4), yielded a good model of the complex (Figure 4.6). Table 4.6 lists the intermolecular residue-residue interactions that make important contributions to the stability of the modeled complex. Our final model is slightly more stable than our initial complex with conformer 1, and this improvement is largely due to interactions involving the Gal*N*Ac moiety (Table 4.6). Hydrogen bonds

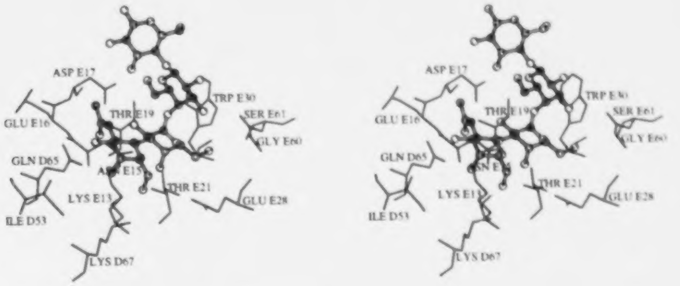

ASP E17
GLU E16
THR E19
TRP E30
SER E61
GLY E60
GLN D65
ASN E15
LYS E13
THR E21
GLU E28
ILE D53
LYS D67

Figure 4.6: The final model of Gb4 bound at site III of SLT-IIe.

donor	acceptor	H· · ·A distance (Å)	D· · ·H· · ·A angle (degrees)
Gal*N*Ac:HO3	GlnD65:OE1	2.36	141.9
LysD67:HZ1	Gal*N*Ac:O6	1.89	161.1
LysE13:HZ1	Gal*N*Ac:O7	2.17	119.1
LysE13:HZ3	Gal*N*Ac:O1	2.29	136.8
GluE16:HN	Gal*N*Ac:O7	2.81	152.6
ThrE21:HG1	Gal2:O4	2.28	168.5
ThrE21:HG1	Gal2:O5	2.42	113.1
Gal2:HO4	GluE28:OE2	2.69	177.8
Gal3:HO6	AspE17:OD2	1.90	150.3
Gal3:HO3	GlyE60:O	2.51	123.6

Table 4.7: Possible intermolecular hydrogen bonds of the final model of the complex of Gb4 bound at site III of SLT-IIe. We used a relatively permissive hydrogen bond filter of 90° for the D· · ·H· · ·A angle, and 3.0 Å for the H· · ·A distance.

between the Gal*N*Ac ring of Gb4 and both GlnD65 and Lys D67 stabilize the complex (Figure 4.6, Tables 4.6, 4.7).

An alternative conformation of the *N*-acetyl group is also possible. Adjustment of the H-N-C-H torsion from approximately 160° (the value in our final model, shown in Figure 4.6) to -140°, the favored value of this torsion for Gb4 in solution (Poppe et al., 1990b), leads to loss of the GalNAc:O7-LysE13 hydrogen bond (Figure 4.6, Table 4.7) and gain of a GalNAc:O7-GlnD65:NE hydrogen bond (not shown). The change in the energetic contribution towards stability of the complex would probably be minimal, but the intramolecular strain energy of Gb4 might be favorably decreased. Also, the newly formed hydrogen bond would increase the apparent importance of Gln65 (see above), and, at least partially, explain the effect of the Gln65→Glu substitution.

4.3.6 Plausibility of the model

We can determine the plausibility of our final model of the Gb4/SLT-IIe complex in several ways. First, we can examine the final conformation of Gb4 itself, and compare it to our calculated low energy conformer(s). Second, we can examine the final protein conformation, to determine if construction of our model has resulted in the generation of any unreasonable structural features. Third, we can compare the Gb3 moiety of our modeled Gb4/SLT-IIe complex to our initial model of Gb3 bound to site III of SLT-IIe. Fourth, and finally, we can examine the consistency of our model with a variety of experimental mutation and binding data.

The final glycoside dihedrals (ϕ/ψ) of our model of bound Gb4 are 32.2/-46.5, -41.9/-8.2, 53.7/0.4 for Gal*N*Acβ1-3Gal, galabiose, and lactose, respectively. These were determined following minimization of the hydrogen atoms in the absence of protein, to allow for a meaningful comparison with our earlier result (100 steps of conjugate gradients minimization with DISCOVER/AMBER, all heavy atoms fixed). The energy of this conformer was -24.5 kcal mol^{-1}, almost identical to that of the original conformer 1 (Table 4.3), despite the changes in ϕ and ψ, -21° and -38°, respectively, of the βGal*N*Ac(1-3)αGal linkage. This confirms our earlier observation (see above) that this disaccharide shows considerable torsional freedom in the immediate vicinity of this energy minimum (Figure 4.1). Of particular interest is the observation that our final Gb4 conformation is much closer than our initial calculated Gb4 conformer to the majority of low energy conformers described in the NMR/computation study of Poppe et al. (1990a). The galabiose and lactose dihedrals of our final Gb4 model (see above) remain essentially unaltered from their initial calculated values (Table 4.3). Overall, therefore, the final modeled conformation of Gb4 bound to SLT-IIe is not much different from our initial calculated low energy Gb4 conformation. Indeed, our manual adjustments led to what is probably a more stable conformation for free Gb4 in solution. It is intriguing to observe that the carbohydrate moieties of both Gb3 and Gb4 bind to these toxins in conformations very similar to those favored by the free carbohydrates in solution.

It is also important to consider whether our model of the Gb4 carbohydrate bound to SLT-IIe is consistent with a Gb4 glycolipid molecule that is a membrane component. The terminal glucose residue of our model is completely exposed to solvent on the membrane binding face of the pentamer. Therefore the glucose residue is "pointing" in a functionally sensible direction, and is unconstrained by any contacts with protein. This is similar to the experimentally observed Gb3 glycolipid analogs bound to GT3 and SLT-I (Ling et al., unpublished coordinates). Furthermore, the flexibility of the βGal(1-4)βGlc disaccharide is well-documented (Poppe et al., 1990a, and refs therein). This, in conjunction with the lack of glucose-protein contacts (noted above), indicates that our model does not exclude any of the glucose-ceramide conformers described by Strömberg et al. (1991) as accessible for the Gb4 glycolipid in a membrane (they used the HSEA-derived low energy conformer of Gb4, a conformation similar to that of our model). Our model of the carbohydrate portion of Gb4 bound to site III of SLT-IIe is therefore consistent with the membrane receptor function of the Gb4 glycolipid.

We used the protein analysis program PROCHECK (Laskowski et al., 1993) to analyze the initial and final model of the SLT-IIe binding subunit trimer used throughout our simulations (results not shown). With the exception of sidechain conformations in the region of the modeled Gb4 binding site, the final and initial SLT-IIe models are virtually identical. Several peptide bonds in this region showed slight increases (1°-3°) in the deviation of ω from ideality, and a few others showed improvements of similar magnitude. Sidechain RMS differences were all less than 1 Å (we ignored the two modified residues). Our manual manipulations of LysD67 and LysE13 (see above) were restricted to the two terminal dihedrals, which are not considered by PROCHECK during sidechain conformation analysis. However, as noted above, these angles were adjusted in 120° steps, and the final conformations seem reasonable (Figure 4.6). All secondary structural features of the initial model were preserved in the final model. Therefore, construction of our model of Gb4 bound to SLT-IIe wrought minimal changes in the protein conformation, and did not result in the generation of any unreasonable features in the conformation of the protein.

Our starting model of Gb3 bound at site III of SLT-IIe was generated by superimposing the SLT-I structure onto our modeled SLT-IIe structure, and then superimposing the galabiose moieties of our calculated Gb4 conformers onto the corresponding atoms of the three Gb3s "bound" to SLT-IIe (see above). For conformer 1 at site III of SLT-IIe, the RMS difference for the Gb3 heavy atoms of the two molecules was 2.0 Å. For the galabiose heavy atoms this difference was 1.6 Å. For our final model of Gb4 bound at this site these differences were 1.8 Å, and 1.1 Å, respectively. The modeling described above has led to a structure with several favorable contacts between Gal*N*Ac and the protein (see above), while at the same time yielding a conformation of the Gb3 moiety which is closer to that observed for Gb3 bound to SLT-I. This consistency with direct experimental observation in a related system supports our model.

The final model of Gb4 bound to SLT-IIe (Figure 4.6) has many features typical of other carbohydrate-protein complexes (Quiocho, 1989; Vyas, 1991). These are elaborated in Tables 4.6 and 4.7. Stacking interactions between carbohydrate rings and aromatic sidechains are a common feature of carbohydrate-protein complexes (Quiocho, 1989; Vyas, 1991). In Table 4.6 we see that Gal2 makes significant contacts with TyrE29 and TrpE30, and that the Gal3-TrpE30 interaction is a major contributor to the stability of the modeled complex. Most of this binding energy derives from van der Waals' interactions (in this case 90%; energy breakdown not shown). The presence of hydrogen bond donating and accepting sidechains is another common feature of carbohydrate binding sites in proteins (Quiocho, 1989; Vyas, 1991). All four Gb4 residues are involved in electrostatic interactions with such sidechains of SLT-IIe (Table 4.6; energy breakdown not shown). Our model includes ten possible hydrogen bonds between the Gal$N$Acβ1-3Galα1-4Gal moiety of Gb4 and SLT-IIe (Table 4.7), and several of these involve acidic or polar sidechains. (We used a relatively permissive hydrogen bond filter to allow for the inexactness of our model.)

Finally, we consider the mutation and binding data specific to the interaction of SLT-IIe with Gb3 and Gb4. The direct involvement of Trp30, the Glu16-Asp17-Asn18 sequence, and Gln65 and Lys67 in complex formation with Gb3/Gb4 has been inferred from the binding changes observed when these residues are altered (discussed above). Our result is consistent with these earlier conclusions, in that together these residues contribute approximately 40% of the total favorable binding energy that stabilizes our modeled complex (Table 4.6). That 45% of the net favorable binding energy derives from interactions involving the Gal$N$Ac residue of Gb4 (Table 4.6) is also compelling, when we consider that this is the moiety that interacts with the two mutated residues of GT3. The possibility that site III is strictly a Gb4 binding site in SLT-IIe and GT3 is consistent with, but not established by, this result. The more recent observation of partial occupancy of this site by Gb3 (see Footnote 3) argues against this conclusion; however the relationship between the crystallographically observed sugar binding sites and the biologically relevant receptor binding sites is not clear.

It is satisfying that Gln65 and Lys67 play an essential role in the specific interaction with the GalNAc moiety of our modeled complex (Tables 4.6 and 4.7, Figure 4.6). The reduced Gb4 affinity of GT3 may be explained by the absence or disturbance of one or both of these interactions in the double mutant (discussed in detail below). Such interactions were first postulated by DeGrandis et al. (1989), and later discussed by Tyrrell et al. (1992) in their original description of this mutant. Our model shows that such interactions are indeed possible for a Gb4 molecule bound at this site of SLT-IIe, and that this explanation of the altered binding activity of GT3 is reasonable (discussed in detail below).

One less than satisfactory aspect of our model is that we do not see any clearly prohibitive interactions between GT3 and Gb4 (not shown). Although GlnD65 and LysD67 make two, or possibly three (see below), hydrogen bonds with the Gal*N*Ac residue of Gb4, their absence does not obviously preclude Gb4 binding at this site. Neither substitution seems to lead to an unfavorable steric or electrostatic interaction (not shown). It may be that the interactions of these two residues with Gal*N*Ac contribute a critical amount of binding energy, and that this critical level is still exceeded when either one of the residues are substituted [either substitution alone does not seem to affect Gb4 binding (Tyrrell et al., 1992)]. The Gln65→Glu substitution would only lead to partial loss of the contribution of this residue, and the remaining hydrogen bond with HO3 of the Gal*N*Ac residue would probably be slightly more favorable with a charged Glu than with Gln. To complete this scenario, then, substitution of both residues would result in loss of a critical amount of binding energy, such that Gb4 binding decreases dramatically (Tyrrell et al., 1992). Of course, it is also possible that binding of Gb3 and/or Gb4 at this site invokes some subtle conformational change(s) that would more clearly explain the altered Gb3 and Gb4 affinities of GT3. There seems to be no clear evidence, however, that strongly supports either of these speculations.

Further inspection of the model (Figure 4.6) suggests that the acetyl group of Gb4 is a relatively minor contributor to complex stability. This is consistent with the slightly decreased affinity observed for de-acetylated Gb4 (DeGrandis et al., 1989), and also applies to the alternative conformation of the *N*-acetyl group discussed above (end of preceding section).

Asn18 is not a major contributor to binding in our model, but this residue may not be as important to the Gb3/Gb4 interaction with SLT-IIe as Asp18 is to the similar interaction with SLT-I (Tyrrell, 1992). Alternatively, the binding change observed (Gb3 only→Gb3 and Gb4) for the Asp18→Asn mutation of SLT-I (Tyrrell et al., 1992) may be due to effects on the specificity of sites I and/or II. The structure of SLT-I with three unique Gb3 binding sites (Ling et al., unpublished coordinates) indicates that direct contacts between the bound carbohydrate and Asp/AsnA18 are more likely at sites I and II than at site III (Figure 4.4). AspA18 forms a salt bridge with ArgE33, which in turn forms a hydrogen bond with O3 of the terminal αGal of Gb3 bound at site I of SLT-I (this aspect will be explored more fully when the structure is reported - Ling et al., manuscript in preparation). A direct hydrogen bond between AspA18 and O4 of the terminal αGal of Gb3 bound at site II is also possible (Figure 4.4). Modeling the AspA16→His and AspA17→His substitutions (not shown) suggests at least two possible explanations for the binding changes observed for this double mutant (Jackson et al., 1990). AspA16→His leads to a loss of a hydrogen bond and gain of a prohibitive steric clash with the terminal αGal moiety of Gb3 bound at site I. Alternatively, one or both of these mutations could lead to local sterically- or electrostatically-driven conformational changes which could disturb either or both sites I and III (not shown, but see Figure 4.4).

The importance of LysE13 to the intermolecular interaction (Tables 4.6 and 4.7, Figure 4.6) suggests that mutagenesis of this residue might provide further information regarding Gb4 binding to SLT-IIe. Also, in single, double, and triple mutants, the Ile53→Lys substitution has implicated Ile53 in receptor binding (Jackson et al., 1990; Tyrrell et al., 1992). However, this issue is clouded by the observation that this substitution (Ile53→Lys) may have major effects on the conformation of the B subunit, as well as on the association of the A and B subunits (Tyrrell et al., 1992). Our model places Gal*N*Ac near IleD53, thus posing a plausible explanation of this result, and also suggesting that more conservative substitution of this residue would be informative.

In conclusion, it is clear that there are compelling reasons to accept the model that we have described. The conformation of the bound Gb4 model is similar to our calculated minimum energy conformation, and is very similar to the reported solution structure of this molecule (Poppe et al., 1990b). The structure of the final protein model is also reasonable. Only relatively minor sidechain adjustments of the SLT-IIe binding site were required to optimize the interaction with Gb4. The relative energetic importance of the contacts between bound Gb4 and the two residues that were mutated in the GT3 double mutant presents a reasonable explanation of the Gb4→Gb3 switch in binding preference exhibited by the double mutant, and strongly supports our model. Several other residues known to be important to Gb3 and/or Gb4 binding are also at or near to the modeled binding site. We look forward to the experimental determination of the structure of the modeled complex.

4.4 References

Armstrong, G. D., Fodor, E., & Vanmaele, R. (1991). Investigation of shiga-like toxin binding to chemically synthesized oligosaccharide sequences. *J. Infect.*

Dis., 164:1160–1167.

Bass, M. B., Hopkins, D. F., Jaquysh, W. A. N., & Ornstein, R. L. (1992). A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. *Proteins: Structure, Function, and Genetics*, 12:266–277.

Boyd, B., Tyrrell, G. J., Maloney, M., Gyles, C., Brunton, J. L., & Lingwood, C. A. (1993). Alteration of the glycolipid binding specificity of the pig edema toxin from globotetraosyl to globotriaosyl ceramide alters in vivo tissue targetting and results in verotoxin 1-like disease in pigs. *J. Exp. Med.*, 177:1745–1753.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248:338–339.

Clark, C., Bast, D., Sharp, A., St. Hilaire, P., Agha, R., Stein, P. E., Toone, E. J., Read, R. J., & Brunton, J. L. (1995). Phenylalanine 30 plays an important role in receptor binding of verotoxin-1. *Mol. Microbiol.*, in press.

Cummings, M. D., Hart, T. N., & Read, R. J. (1995a). Atomic solvation parameters in the analysis of protein-protein docking results. *Protein Sci.*, 4:2087–2099.

Cummings, M. D., Hart, T. N., & Read, R. J. (1995b). Monte Carlo docking with ubiquitin. *Protein Sci.*, 4:885–899.

DeGrandis, S., Law, H., Brunton, J., Gyles, C., & Lingwood, C. A. (1989). Globotetraosylceramide is recognized by the pig edema disease toxin. *J. Biol. Chem.*, 264:12520–12525.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, 319:199–203.

Eisenberg, D., Wesson, M., & Yamashita, M. (1989). Interpretation of protein folding and binding with atomic solvation parameters. *Chemica Scripta*, 29A:217–221.

French, A. D. & Brady, J. W. (1990). Computer modeling of carbohydrates: an introduction. In French, A. & Brady, J., editors, *Computer modeling of carbohydrate molecules*, pages 1–19. American Chemical Society, Washington, DC.

Hart, T. N. & Read, R. J. (1992). A multiple-start Monte Carlo docking method. *Proteins: Structure, Function, and Genetics*, 13:206–222.

Homans, S. W. (1990). A molecular mechanical force field for the conformational analysis of oligosaccharides: comparison of theoretical and crystal structures of Manα1-3Manβ1-4Glc*N*Ac. *Biochemistry*, 29:9110–9118.

Jackson, M. P., Wadolkowski, E. A., Weinstein, D. L., Holmes, R. K., & O'Brien, A. D. (1990). Functional analysis of the shiga toxin and shiga-like toxin type II variant binding subunits by using site-directed mutagenesis. *J. Bacteriol.*, 172:653–658.

Kraulis, P. J. (1991). Molscript: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, 24:946–950.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, 26:283–291.

Ling, H., Boodhoo, A., Armstrong, G. D., Brunton, J. L., & Read, R. J. (1995). Shiga-like toxin I B-subunit complexed with a cell-surface trisaccharide. In ACA, editor, *Proceedings of the ACA meeting in Montreal*. Academic Press, San Diego. W127.

Lingwood, C. A. (1993). Verotoxins and their glycolipid receptors. In Bell, R., Hannun, Y., & Merrill, A., editors, *Advances in lipid research*, pages 189–212. Academic Press, San Diego.

Merritt, E. A., Sarfaty, S., van den Akker, F., L'hoir, C., Martial, J. A., & Hol, W. G. J. (1994a). Crystal structure of cholera toxin B-pentamer bound to receptor G_M1 pentasaccharide. *Protein Sci.*, 3:166–175.

Merritt, E. A., Sixma, T. K., Kalk, K. H., van Zanten, B. A. M., & Hol, W. G. J. (1994b). Galactose-binding site in *Escherichia coli* heat-labile enterotoxin (LT) and cholera toxin (CT). *Mol. Microbiol.*, 13:745–753.

Nyholm, P.-G., Brunton, J. L., & Lingwood, C. A. (1995). Modelling of the interaction of verotoxin-1 (VT1) with its glycolipid receptor, globotriaosylceramide (Gb_3). *Int. J. Biol. Macromol.*, 17:199–204.

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different classes. *J. Mol. Biol.*, 193:775–791.

Poppe, L., Dabrowski, J., von der Lieth, C.-W., Koike, K., & Ogawa, T. (1990a). Three-dimensional structure of the oligosaccharide terminus of globotriaosylceramide and isoglobotriaosylceramide in solution. *Eur. J. Biochem.*, 189:313–325.

Poppe, L., von der Lieth, C.-W., & Dabrowski, J. (1990b). Conformation of the glycolipid globoside head group in various solvents and in the micelle-bound state. *J. Am. Chem. Soc.*, 112:7762–7771.

Quiocho, F. A. (1989). Protein-carbohydrate interactions: basic molecular features. *Pure. Appl. Chem.*, 61:1293–1306.

Sixma, T. K., Stein, P. E., Hol, W. G. J., & Read, R. J. (1993). Comparison of the B-pentamers of heat-labile enterotoxin and verotoxin-1 - 2 structures with remarkable similarity and dissimilarity. *Biochemistry*, 32:191–198.

St. Hilaire, P. M., Boyd, M. K., & Toone, E. J. (1994). Interaction of the shiga-like toxin type 1 B-subunit with its carbohydrate receptor. *Biochemistry*, 33:14452–14463.

Stein, P. E., Boodhoo, A., Tyrrell, G. J., Brunton, J. L., & Read, R. J. (1992). Crystal structure of the cell-binding B oligomer of verotoxin-1 from *E. coli*. *Nature*, 355:748–750.

Stuike-Prill, R. & Meyer, B. (1990). A new force-field program for the calculation of glycopeptides and its application to a heptacosapeptide-decasaccharide of immunoglobulin G_1. *Eur. J. Biochem.*, 194:903–918.

Tvaroška, I. & Pérez, S. (1986). Conformational energy calculations for oligosaccharides: a comparison of methods and a strategy for calculation. *Carbohydr. Res.*, 149:389–410.

Tyrrell, G. J., Ramotar, K., Toye, B., Boyd, B., Lingwood, C. A., & Brunton, J. L. (1992). Alteration of the carbohydrate binding specificity of verotoxins from Galα1-4Gal to Gal$N$Acβ1-3Galα1-4Gal and vice versa by site-directed mutagenesis of the binding subunit. *Proc. Natl. Acad. Sci. USA*, 89:524–528.

Vyas, N. K. (1991). Atomic features of protein-carbohydrate interactions. *Curr. Opin. Struct. Biol.*, 1:732–740.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. J., & Weiner, P. (1984). A new force-field program for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784.

Chapter 5

Fragment-based modeling of NAD binding to the catalytic subunits of diphtheria and pertussis toxins[1]

5.1 Introduction

One approach to the structure-based design of ligands specific for the binding site of a target protein is known as *fragment-based* ligand (or drug) design. Given a binding site of interest, docking simulations are performed with small molecular fragments (*eg.* functional groups, heterocycles). Favorable dockings of each fragment are collected, and potential ligands are constructed by connecting the various docked fragments. Novel molecules may be suggested, and a large variety can be generated from relatively few simple fragments; in many ways this simulation method is analogous to the *real* process of combinatorial chemistry. Many such computational methods have been reported (*eg.* (DesJarlais et al., 1986; Moon & Howe, 1991; Bohm, 1992; Rotstein & Murcko, 1993)). The approach most similar to that of the present work was the "binary docking" method described by Stoddard & Koshland (1992). Here we describe a novel application of the method. Using the structure of nicotinamide adenine dinucleotide (NAD) bound to the catalytic subunit of diphtheria toxin (DT), we show that docking simulations involving the NAD fragments nicotinamide and adenine and the catalytic subunits of both DT and pertussis toxin (PT) provide information relevant to the prediction of a reasonable model of NAD binding to PT. Thus, given one or more related protein-ligand structures, fragment docking can be useful in predicting the structures of other related complexes. Our results also serve to further validate the idea that fragment docking is a useful method for ligand design.

---

[1] A version of this chapter has been submitted for publication: M.D. Cummings, T.N. Hart, & R.J. Read, *Fragment-based modeling of NAD binding to the catalytic subunits of diphtheria and pertussis toxins*.

DT belongs to the AB class of ADP-ribosyltransferase toxins (ADPR toxins; reviewed in (Read & Stein, 1993; Merritt & Hol, 1995)) that includes PT, *Pseudomonas aeruginosa* exotoxin A (ETA), the heat-labile enterotoxin of *E. coli* (LT), and cholera toxin (CT). These toxins evoke their toxic response after releasing a catalytic ADP-ribosyltransferase subunit in the cytoplasm of a host cell. Intracellular ADP-ribosylation of specific targets by the catalytic subunit leads to disturbance of host cell metabolism [reviewed in (Moss & Vaughan, 1988)]. The catalytic subunits of these toxins bind NAD, and exhibit *in vitro* NAD-glycohydrolase and ADP-ribosyltransferase activity.

Unfortunately, little is known of the relevant mechanistic details, although much has been inferred. For example, although it seems likely that the catalytic mechanisms of the various ADPR toxins are similar [eg. (Domenighini et al., 1991; Domenighini et al., 1994)], it has not been clearly established that this is the case. Similarly, it is not clear for any toxin whether the reaction proceeds *via* an S_N1 or S_N2 pathway [eg. (Wilson et al., 1990)], although an S_N2 route was suggested by one mechanistic study with CT (Soman et al., 1986), and has been inferred from the stereochemical inversion observed for the initial reaction products obtained with CT (Oppenheimer, 1978) and LT (Moss et al., 1979). Key residues for NAD binding and ADP-ribosyltransferase activity have been identified for several of the toxins [reviewed in (Domenighini et al., 1994)]. PT and LT ADP-ribosylate a Cys and Arg residue, respectively, of various GTP-binding proteins involved in signal transduction; DT and ETA act on a diphthamide residue of elongation factor 2 (EF-2). These similarities and differences are reflected in the observed sequence and structural relationships of the toxins: LT and PT are most similar to each other, and DT and ETA form another similar pair. Since the catalytic mechanisms are unknown, precise functions cannot be ascribed to specific residues with confidence.

The structure of adenylyl 3'-5' uridine 3' monophosphate bound to DT [ApUp-DT; (Bennett & Eisenberg, 1994)] and, more recently, those of the NAD-DT (Bell & Eisenberg, 1996) and hydrolyzed NAD-ETA [hNAD-ETA; (Li et al., 1995)] complexes, have confirmed many speculations regarding the NAD binding sites of DT and ETA, including the proximity of many catalytically important residues to the bound ligands. Structural comparison of the proposed NAD binding sites of several toxins (Stein et al., 1994; Domenighini et al., 1994) showed that the spatial relationship of these key residues is largely conserved in the different binding sites. Overall, ADPR toxins exhibit low sequence identity (Read & Stein, 1993; Domenighini et al., 1994; Stein et al., 1994; Merritt & Hol, 1995); however, the NAD binding sites are structurally conserved (Domenighini et al., 1994; Li et al., 1995; Bell & Eisenberg, 1996). On this basis, and given the NAD-DT and hNAD-ETA structures, the prediction of the binding mode of NAD to the related toxins for which structures are available seemed feasible. Indeed, we expected this to be a relatively trivial problem, given the structural and functional similarity of the toxins, as well as the virtual identity of the NAD-DT and hNAD-ETA structures. However, we did not find this to be the case. In their description of the NAD-DT complex, Bell & Eisenberg (1996) discussed, in general terms, the predicted NAD binding sites for several related toxins, including PT. Here we present a detailed modeling study of the NAD-PT complex, and show that while NAD binding to DT and ETA is very similar, there appear to be significant differences between NAD binding to PT and to either DT or ETA.

5.2 Materials and methods

5.2.1 Preparation of docking targets and ligands

The structure of the catalytic portion of DT used as a docking target comprised residues 1-187 of 1mdt [(Bennett & Eisenberg, 1994); note that this is the structure of DT with ApUp bound, *not* NAD]. The PT structure used as a docking target included residues 2-180 of the catalytic subunit of 1prt (Stein et al., 1994). This truncated form of the S1 subunit of PT retains full NAD binding and hydrolytic activity, but has reduced ADPR activity (Cortina & Barbieri, 1991). Residues 199-207 form a helix that occupies the NAD binding site of the unactivated toxin (Stein et al., 1994; Bell & Eisenberg, 1996), so their removal was essential to this modeling study. Both of these structures, as well as that of hNAD-ETA [1dma; (Li et al., 1995)], were obtained from the Brookhaven Protein Data Bank (Bernstein et al., 1977). Adenine and nicotinamide were taken from the templates provided in INSIGHTII (Biosym Technologies Limited, San Diego), and NAD was from the NAD-DT structure (Bell & Eisenberg, 1996). Neutral charge groups were used for all docking simulations, as described previously (Hart & Read, 1992). General structure manipulation and visualization were performed with INSIGHTII, DISCOVER, and other modules of Biosym software (Biosym Technologies Limited, San Diego).

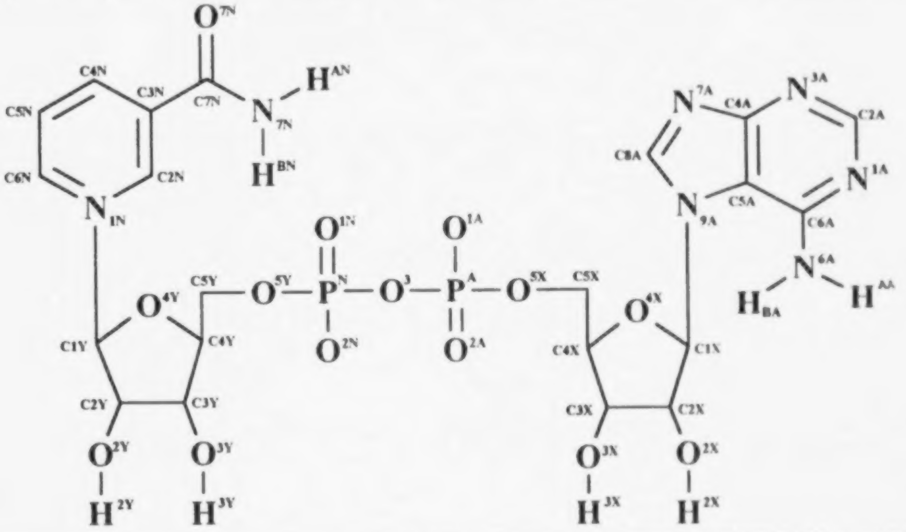
	topology file^b				
dihedral^a	1	2	3	4	5
C1Y-C2Y-O2Y-H2Y	-	-	-	+	+
C2Y-C3Y-O3Y-H3Y	-	-	-	+	+
C1X-C2X-O2X-H2X	-	-	-	+	+
C2X-C3X-O3X-H3X	-	-	-	+	+
O4Y-C1Y-N1N-C2N	-	-	+	+	-
C3Y-C4Y-C5Y-O5Y	+	-	+	+	+
C4Y-C5Y-O5Y-PN	-	+	+	+	+
C5Y-O5Y-PN-O3	+	-	+	+	+
O5Y-PN-O3-PA	-	+	+	+	+
PN-O3-PA-O5X	+	-	+	+	+
O3-PA-O5X-C5X	-	+	+	+	+
PA-O5X-C5X-C4X	+	-	+	+	+
O5X-C5X-C4X-C3X	-	+	+	+	+
O4X-C1X-N9A-C4A	+	-	+	+	-

Table 5.1: Flexible NAD dihedrals for superposition and docking/refinement. ^{a}Atom names are shown in Figure 5.1. ^{b}Flexibility allowed (+) or not allowed (-) for this dihedral in this topology file. For the Monte Carlo search during flexible superposition (see Methods), flexible dihedrals were searched in 60° steps. During flexible Monte Carlo docking/refinement (see Methods), flexibility was turned on for *all* fourteen dihedrals listed here, as in topology file 4, and the Monte Carlo minimization schedule was as follows: $kT = 10^{-4}$ kcal mol^{-1}, 3 cycles of 500 steps each with maximum rotations (dihedral or rigid-body)/maximum translations of 3°/0.1Å, 2°/0.08Å, and 1°/0.04Å, respectively.

Hydrogens were added to all the appropriate heavy atoms of the target proteins. Polar hydrogen positions were first optimized with NETWORK (Bass et al., 1992); all hydrogens were then minimized (heavy atoms fixed) for 200 cycles of steepest descents minimization followed by a maximum of 200 cycles of conjugate gradient minimization with the CVFF potential function of the DISCOVER program (Biosym Technologies Limited, San Diego). Mobile polar hydrogens of NAD were not optimized during structure preparation, since the relevant torsional flexibility was considered during superpositioning and Monte Carlo refinement (see below).

5.2.2 Superpositioning and rigid-body energy minimization

Rigid and flexible superpositioning and rigid-body conjugate gradient minimization of probe-target configurations were performed with unpublished programs developed by Trevor Hart. We used flexible superpositioning to generate different initial NAD conformers. In this procedure (to be described in detail elsewhere) the adenine, nicotinamide, and ribose rings of NAD were fixed in their initial conformations; all bonds connecting these fragments, as well as the four hydroxyls of the ribose rings, were defined as dihedrals in a topology file. Five different topology files were created, in which different dihedrals were arbitrarily fixed or free to rotate in 60° steps (Figure 5.1, Table 5.1). Flexible superpositioning starts with a least-squares rigid superpositioning of the adenine and nicotinamide moieties of NAD onto the chosen adenine/nicotinamide pairs of dockings (see below). This is followed by a Monte Carlo search of the flexible dihedrals specified in the topology file, while minimizing the root-mean-square (RMS) distance between the selected ring atoms of the NAD molecule and the docked fragment pairs. A variety of different topology files (Table 5.1) thus yields a range of different NAD conformers from this procedure, by restricting the different searches to different regions of conformational space. The



O7N
C4N
C3N
C7N
N7N
HAN
HBN
C5N
C6N
C2N
N1N
O1N
O1A
C5Y
O5Y
PN
O3
PA
O5X
C5X
O4Y
O4X
C1Y
C4Y
O2N
O2A
C4X
C1X
C2Y
C3Y
C3X
C2X
O2Y
O3Y
O3X
O2X
H2Y
H3Y
H3X
H2X
N7A
C4A
N3A
C2A
C8A
N1A
N9A
C5A
C6A
N6A
HBA
HAA

Figure 5.1: NAD atom names.

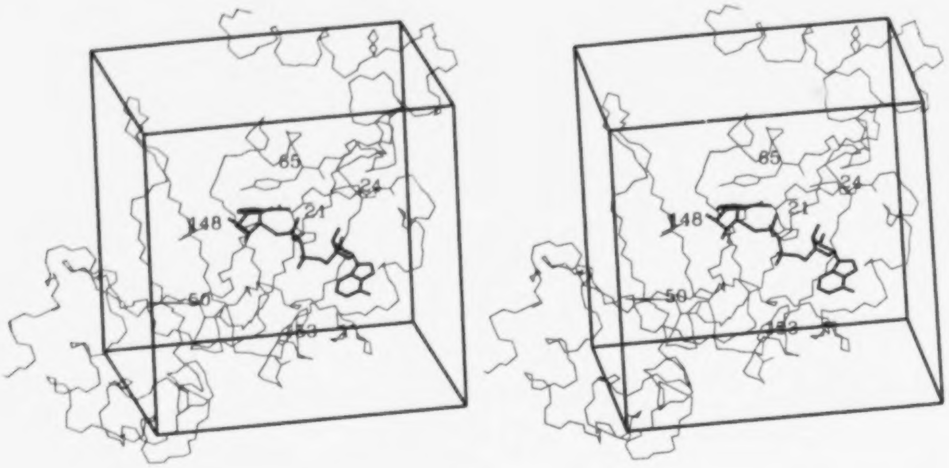


Figure 5.2: Extents of the fragment docking searches. The backbone of the DT docking target is shown (protein was taken from the ApUp-DT structure) with NAD bound (from superposition) and selected important sidechains highlighted. The docking space was a box 30Å on each side.

conformational search includes an internal check for atomic overlap, but does not consider any interactions with the target protein. Various initial NAD conformers were used.

5.2.3 Docking simulations

A rigid-body multiple start Monte Carlo docking method (BOXSEARCH) was used for docking nicotinamide and adenine to DT and PT. This method has been described in detail elsewhere (Hart & Read, 1992; Hart & Read, 1994; Cummings et al., 1995). Each docking simulation consisted of 20,000 separate runs. A BOXSEARCH run begins with the random placement of the probe in the box (Figure 5.2) which defines the limits of the search space. The probe is then quickly "floated" to the surface of the target protein, and a simulated annealing schedule is then invoked. If the interaction energy of the complex is below the user-selected energy cutoff at the end of this schedule, further low temperature simulated annealing is performed and the docking is written to output. All of the output dockings are then further refined with a rigid-body conjugate gradient minimizer (unpublished program of Trevor Hart). Finally, the refined output list is divided into "clusters" of similar dockings, and the lowest energy member of each cluster is saved (Hart & Read, 1992).

We selected adenine/nicotinamide pairs of the dockings obtained by the procedure described above. Using the flexible superposition method described above, we generated twenty initial dockings of NAD to PT, by superimposing NAD onto the selected adenine/nicotinamide pairs. These dockings were further refined in two stages. We began with a new Monte Carlo-simulated annealing flexible docking method (to be described in detail elsewhere), which represents a significant extension of the original BOXSEARCH docking method. In the present study we have used this new docking method in a very limited way, to refine what were already reasonable dockings. The geometric centre of mass of the NAD docking was restricted to remain within a sphere 10Å in diameter, the fourteen flexible dihedrals described above were *all* allowed free rotation in relatively small steps (Table 5.1), and a relatively low temperature Monte Carlo simulation was performed. Simulated annealing was not used in this refinement procedure.

```
    7      23   35  50            84         98 128      140
PT  VYRYDS  FTA  HL  FVSTSSSRRYTE  FIGYIYEVR  YG  SEYLAHR  NIRRV
    5      22   44  59            81         95 111      123
LT  LYRADS  LMP  HA  YVSTSLSLRSAH  STYYIYVIA  FN  QEVSALG  QIYGW
  438     455      468           494        541 552      565
ETA GYHGTF  VRA  --  GFYIAGDPALAY  RNGALLRVY  AI  LETILGW  VVIPS
   19      35       52            76        136 147      160
DT  SYHGTK  IQK  --  GFYSTDNKYDAA  KAGGVVKVT  LS  VEYINNW  SVELE
```

Figure 5.3: Sequence alignment used for superpositioning of DT, PT, and ETA (taken from Stein *et al.*, 1994).

Following the initial stage of refinement, two models of the NAD-PT complex were more extensively refined with the AMBER potential function (Weiner et al., 1984),

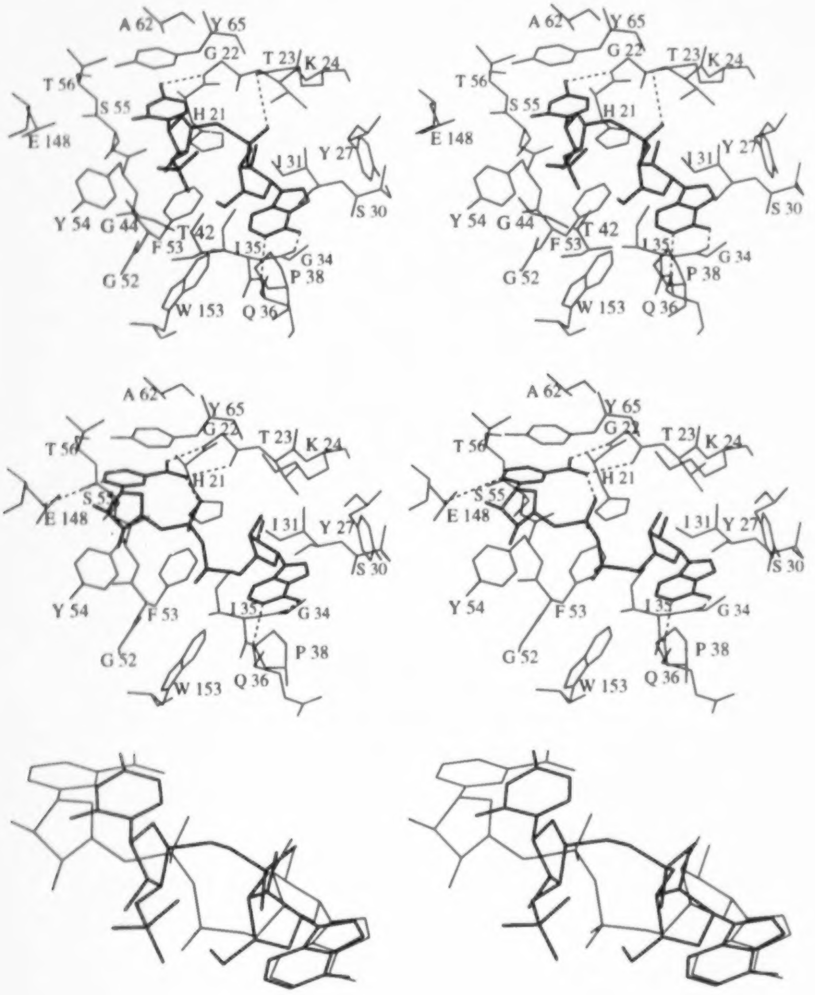

A 62
Y 65
G 22
T 23
K 24
T 56
S 55
H 21
E 148
I 31
Y 27
S 30
Y 54
G 44
T 42
F 53
I 35
G 34
G 52
P 38
W 153
Q 36
A 62
Y 65
G 22
T 23
K 24
T 56
S 55
H 21
E 148
I 31
Y 27
S 30
Y 54
F 53
I 35
G 34
G 52
P 38
Q 36
W 153

Figure 5.4: The NAD and ApUp binding site of DT. For the two complexes, most of the residues with an atom within 8Å of a ligand atom are shown. Important contacts referred to in the text are shown with broken lines. For this figure, we superimposed all $C\alpha$s of the proteins in the two complexes. ***Top:*** The ApUp binding site of the ApUp-DT complex; ***middle:*** the NAD binding site of the NAD-DT complex; ***bottom:*** superposition of the ligands from the NAD-DT and ApUp-DT structures.

as implemented in the DISCOVER program (Biosym Technologies Limited, San Diego). Minimization was carried out in three stages. First, all atoms were fixed, except for the hydrogens of NAD and certain critical residues. Second, all hydrogens were unconstrained, as were the heavy atoms of the sidechains of Gln127 and Glu129. In the final round all hydrogens, and all sidechains within 5Å of NAD, were unconstrained (note that the heavy atoms of NAD were constrained throughout this procedure).

5.3 Results and discussion

The main goal of this study was the prediction of the NAD-PT complex. However, there is more structural information available regarding NAD binding to DT (the NAD-DT and ApUp-DT structures) and ETA (the hNAD-ETA structure) than to PT (the *apo*-PT structure only). Since DT and ETA are more closely related to each other than either is to PT, we begin with a brief comparative analysis of the available relevant complex structures (NAD-DT, ApUp-DT, and hNAD-ETA). This is followed by a description of the docking of the adenine and nicotinamide "fragments" to DT and PT, and a comparison of the results obtained with DT to the NAD-DT structure. In the third and final section below, we compare the NAD-DT and *apo*-PT structures, as well as the dockings to these two proteins. We extrapolate a model of the NAD-PT complex from the quite obvious relationship between the dockings to DT and the NAD-DT structure, the fragment dockings to PT, and the structural relationship between DT and PT. Finally, we conclude with a brief consideration of the relevance of our results and analyses to inhibitor design for the ADPR toxins.

5.3.1 Preliminary analysis of relevant available complex structures

Using a previously published structural alignment of DT, ETA, PT, and LT [Figure 5.3; (Stein et al., 1994)], we superimposed the NAD-DT and hNAD-ETA structures onto ApUp-DT. We focus on the adenine and nicotinamide/uracil moieties

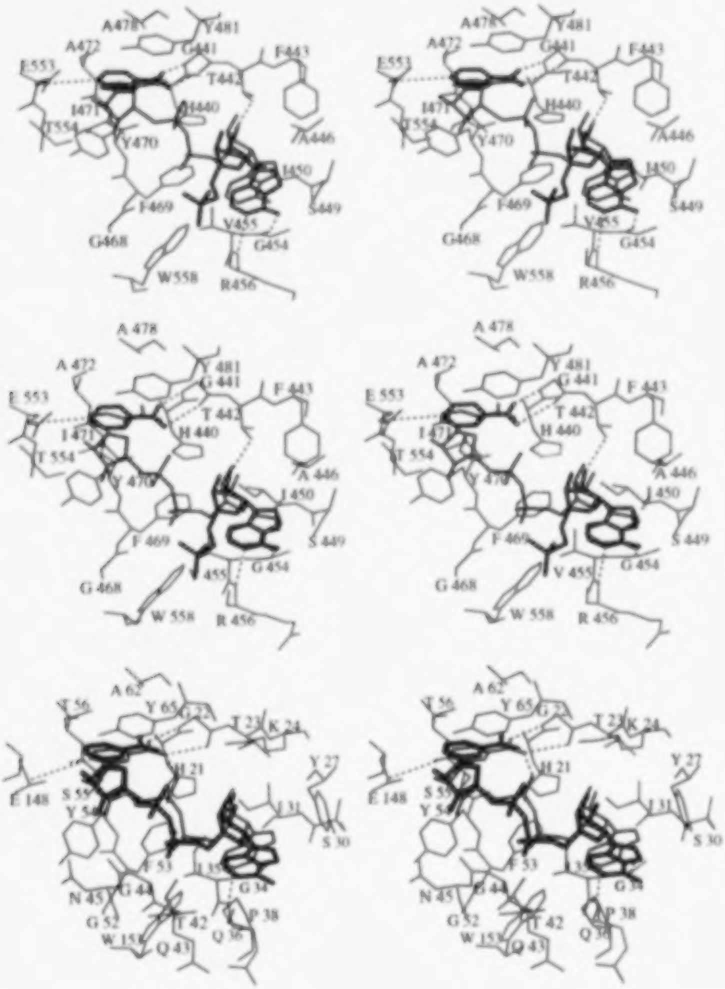

A478
Y481
A472
G441
F443
E553
T442
I471
H440
T554
Y470
A446
I450
S449
F469
V455
G454
G468
W558
R456
A 478
A 472
Y 481
G 441
F 443
E 553
T 442
I 471
H 440
T 554
Y 470
A 446
I 450
F 469
S 449
V 455
G 454
G 468
W 558
R 456
A 62
Y 56
Y 65
G 22
T 23
K 24
Y 27
E 148
H 21
S 55
Y 54
I 31
S 30
F 53
I 35
N 45
G 44
G 34
G 52
T 42
P 38
W 153
Q 43
Q 36

Figure 5.5: The NAD binding site of ETA. Most of the residues with an atom within 8Å of a ligand atom are shown. Important contacts referred to in the text are shown with broken lines. For this figure, we used the superposition described in Figure 5.3. *Top:* The NAD binding site of the hNAD-DT complex. The hydrolyzed NAD is shown in thick lines, with the NAD molecule from the NAD-DT structure superimposed (medium lines; see Figure 5.3). *Middle:* NAD docking to ETA. Same as the top figure, except that the lowest energy NAD docking (see text) to ETA is shown (medium lines), along with the hydrolyzed NAD. *Bottom:* Superposition of the lowest energy NAD docking to ETA (from the middle figure; thick lines) onto the NAD-DT reference structure (NAD from NAD-DT structure, DT from ApUp-DT structure).

of the bound ligands. This preliminary analysis highlights two similarities of direct relevance to the main objective of this work: the similarity of the adenine-toxin binding mode in all three structures, and the similarity of the nicotinamide-toxin binding mode in the NAD-DT and hNAD-ETA structures.

Figure 5.4 compares the binding modes of NAD and ApUp to DT. A subset of DT residues was defined that included all residues with one or more atoms within 5Å of any atom of bound NAD. The RMS difference between all heavy atoms of this subset was 0.7Å; for the adenine moieties of ApUp and NAD bound to DT this difference was 0.4Å. Virtually all adenine-DT contacts are conserved between the two structures. The differences between the remaining (non-adenine) ligand-DT contacts are much greater, reflecting the structural differences between the two dinucleotides [Figure 5.4; see also (Bell & Eisenberg, 1996)]. One nicotinamide-DT hydrogen bond, Gly22:N-O7N (2.8Å), is analogous to a hydrogen bond in the ApUp-DT structure (Gly22:N-O4U; 3.1Å).

Superposition of the NAD-DT and hNAD-ETA structures shows that the interactions between the two proteins and both adenine and (especially) nicotinamide are remarkably similar (Figure 5.5). The differences are somewhat greater between the respective adenine (1.0Å RMS) and nicotinamide (0.7Å RMS) groups than between the adenines of ApUp and NAD bound to DT (see above); this is also true for the residues comprising the binding site. The heavy atom RMS difference between ten strictly conserved residues in the NAD binding sites of the two proteins is 1.2Å. Although the NAD binding sites of DT and ETA are highly structurally conserved, overall the catalytic subunits of DT and ETA show only 23% sequence identity (Carroll and Collier,1988); the similarity between the binding modes of the adenine and nicotinamide moieties to these two proteins is therefore striking, and is strongly suggestive of very similar NAD binding modes to these two enzymes. Since our primary objective was the more challenging prediction of the related NAD-PT structure, we decided to model the NAD-ETA complex as a test case.

We made the reasonable assumption that the observed nicotinamide and adenine binding modes to ETA (in the hNAD-ETA structure) would also apply to the NAD-ETA complex. Prediction of the NAD-ETA complex was, therefore, a relatively simple and straightforward problem. We used flexible superposition with different topology files to generate five initial dockings of NAD to ETA (see Methods). Each of these dockings was used as a starting point for further flexible Monte Carlo refinement (as described above for DT), thus generating a total of 100 NAD-ETA dockings. All 100 dockings were within 2Å RMS of the lowest energy docking. Figure 5.5 shows the lowest energy docking thus obtained, as well as the hydrolyzed NAD fragments bound to ETA.

The bound NAD molecules of the NAD-DT structure and our NAD-ETA model are virtually superimposable. Considering the foundation of this model (the *known* NAD-DT and hNAD-ETA structures), our result strongly supports the proposal that the binding modes of NAD to DT and ETA are similar. Furthermore, this result encouraged us to pursue the more difficult problem of predicting the related structure of the NAD-PT complex.

5.3.2 NAD fragment docking to DT and PT

During initial docking studies with NAD and both DT and PT, we observed a striking result. The low energy dockings from simulations with the rigid nicotinamide and adenine fragments clustered into three distinct regions of both DT and PT. The docking studies with DT involved the protein from the ApUp-DT structure, so it was not too surprising to observe docking of the nicotinamide and adenine heterocycles to regions of the protein that were, presumably, optimized to bind such ring systems. However, the PT used in our docking simulations was from an uncomplexed structure (Stein et al., 1994), and therefore should not be as biased as the DT structure towards planar heterocycles. There is a clear correspondence between the adenine and nicotinamide dockings to DT and the disposition of these moieties in the crystallographic NAD-DT structure. From the structural homology of NAD binding sites of DT and PT, and the similarity between the nicotinamide and adenine dockings to these two proteins, we inferred that these fragment docking results had direct relevance to the structure of the NAD-PT complex. Since the NAD-DT structure provides a known reference of a structure related to the NAD-PT complex, we briefly compare the NAD-DT structure and the fragment dockings to DT.

We docked nicotinamide and adenine to a large region of the DT surface; this was followed by rigid-body conjugate gradient minimization and subsequent cluster analysis (see Methods). To compare these dockings to the observed mode of NAD binding to DT we superimposed the NAD-DT structure onto the DT docking target and determined the RMS differences between the heavy atoms of the fragment dockings and the related atoms of the superimposed NAD molecule (Table 5.2). The lowest energy nicotinamide docking is closest to the nicotinamide ring of bound NAD. Although several of the other low energy dockings are relatively close to the nicotinamide moiety of bound NAD, none are within 2.8Å RMS, with the exception of docking 1. In contrast, not one of the ten lowest energy adenine dockings are within 10Å RMS of the adenine moiety of bound NAD (Table 5.2). Dockings 11 and 13 are relatively close, and docking 18 represents the nearest docking (Table 5.2).

Figure 5.6 shows the DT docking target (from the ApUp-DT structure), the low energy fragment dockings clearly grouped into three discrete regions, and NAD from the superposition of the NAD-DT structure onto the DT target. The sidechains of residues known to be crucial for NAD binding and/or catalytic activity are highlighted. The low energy dockings clearly divide into three discrete clusters (Figure 5.6). Photolabeling of DT with NAD leads to formation of a covalent bond between CG of decarboxylated Glu148 and C6 of the nicotinamide ring [C6N;

cluster	adenine			nicotinamide			
	energy[a] (kcal mol^{-1})	RMS[b] (Å)	DT region[c]	energy[a] (kcal mol^{-1})	RMS[b] (Å)	photolabeling distance[d] (Å)	DT region[c]
1	-31.4	11.3	I	-30.5	0.8	3.8	I
2	-30.0	12.0	I	-30.3	14.1	15.4	IV
3	-28.7	12.1	I	-29.8	3.6	8.4	I
4	-28.6	11.4	I	-29.7	13.6	14.4	IV
5	-28.5	10.9	I	-29.5	14.8	16.4	IV
6	-27.7	12.0	I	-29.3	13.4	14.0	IV
7	-27.7	12.0	I	-29.0	3.4	9.5	I
8	-26.1	14.9	III	-28.1	2.9	6.5	I
9	-25.7	13.6	III	-27.4	2.8	7.4	I
10	-25.3	14.6	III	-26.6	13.6	14.9	IV
11	-25.1	2.8	II	-26.4	3.3	7.0	I
12	-24.3	14.4	III	-26.2	16.4	15.2	III
13	-24.2	2.9	II	-26.2	12.4	14.7	II
14	-23.9	13.8	III	-26.0	16.8	16.0	III
15	-23.8	15.4	III	-24.9	16.4	16.9	III
16	-23.5	14.4	III	-24.9	17.1	14.5	III
17	-23.3	24.4	-	-24.7	20.9	26.6	-
18	-23.3	0.5	II	-24.4	17.4	17.9	III
19	-23.1	3.7	II	-24.4	16.6	16.0	III
20	-22.7	15.2	III	-24.0	14.4	13.0	IV

Table 5.2: Comparison of low energy nicotinamide and adenine dockings with the crystallographic NAD-DT structure. The twenty lowest energy clusters are shown for the adenine (total clusters: 132) and nicotinamide (total clusters: 139) dockings. [a]The calculated interaction energy of the lowest energy member of each cluster of dockings. Following multiple-start rigid-body Monte Carlo docking with BOXSEARCH, each docking was energy minimized with a rigid-body conjugate gradient method prior to cluster analysis (see Methods). [b]Root-mean-square distance between the heavy atoms of that docking and the corresponding atoms of the NAD-DT complex, following superposition onto the DT docking target (see Methods). [c]The low energy dockings fall into one of three discrete regions of the DT surface, as shown in Figure 5.6 (docking 17 is the sole exception for both fragments - see Figure 5.6). [d]DT is photolabelled upon irradiation of the NAD-DT complex. The product thus formed is covalently linked between CG of (decarboxylated) Glu148 and C6 of the nicotinamide ring (Carroll *et al.*, 1985), thus the distance between these two atoms may serve as a distance criterion for analysis of nicotinamide dockings. The distance between these two atoms in the NAD-DT complex is 4.1Å.

(Carroll et al., 1985)]. On the basis of proximity (the Glu148:CG-C6N distance is 4.1Å in the NAD-DT complex), formation of this covalent adduct seems most likely for nicotinamide docking 1 (Table 5.2), and other dockings at region I of DT (Table 5.2, Figure 5.6). All of the dockings in regions II-IV are relatively distant from the catalytically essential residue Glu148 (Table 5.2, Figure 5.6); for these dockings the photolabeling reaction seems unlikely. With this constraint and the structure of NAD

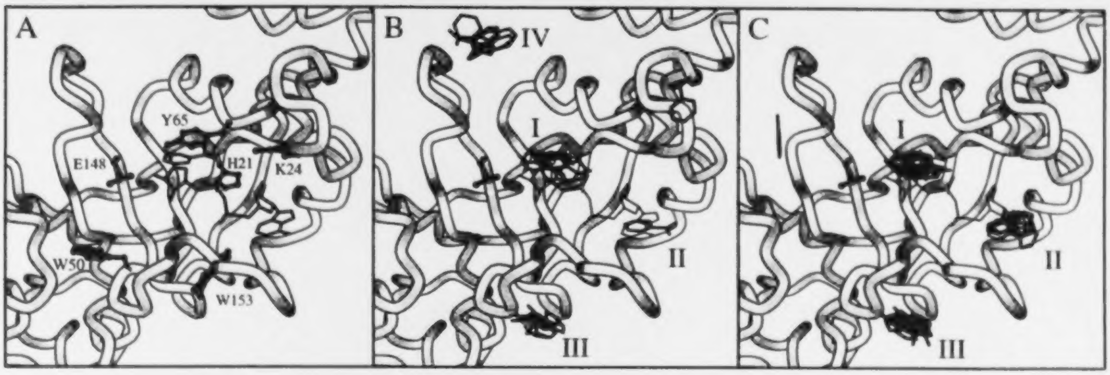


Figure 5.6: Low energy fragment dockings to DT. DT is shown as a ribbon model, with important sidechains displayed in stick form **A**, and bound NAD from superposition of the NAD-DT structure. Glu148 has a catalytic role (Carroll et al., 1985; Wilson et al., 1990), whereas His21 (Papini et al., 1991; Johnson & Nicholls, 1994; Blanke et al., 1994b), Lys24 (Bennett et al., 1994), Trp50 (Michel & Dirkx, 1977; Wilson et al., 1994), Tyr65 (Brandhuber et al., 1988; Papini et al., 1991; Blanke et al., 1994a), and Trp153 (Michel & Dirkx, 1977; Wilson et al., 1994) appear to be involved in substrate binding. The low energy representatives of the twenty lowest energy docking clusters (**B:** nicotinamide; **C:** adenine) are shown, along with bound NAD as in **A**, making the relationship between the docking clusters and bound NAD obvious.

---

in mind, inspection of the adenine dockings would lead one to focus on dockings of this fragment at region II of DT (Figure 5.6). Our results modeling the NAD-ETA complex (see above) establish that, with our flexible superposition tool and these selected fragment dockings to DT, we would arrive at one or a few models of the NAD-DT complex that closely resemble the known NAD-DT structure.

Our fragment docking results suggest that free adenine may bind to the nicotinamide subsite of the observed NAD binding site (Table 5.2, Figure 5.6), and this relates to two earlier experimental results. First, in photolabelling studies with 8-azidoadenine and 8-azidoadenosine (Papini et al., 1991), it was shown that upon photoactivation these compounds react with Tyr65. Papini and co-workers concluded that "nicotinamide and adenine share the same or closely spaced binding sites." Our fragment docking results support this conclusion. Given the distance between adenine dockings at region II (the adenine subsite for NAD binding), labeling of Tyr65 by

cluster	adenine			nicotinamide			
	energy[a] (kcal mol^{-1})	RMS[b] (Å)	PT region[c]	energy[a] (kcal mol^{-1})	RMS[b] (Å)	photolabeling distance[d] (Å)	PT region[c]
1	-27.6	11.6	I	-27.4	4.1	8.1	I
2	-26.7	6.2	II	**-26.6**	**0.7**	**4.5**	**I**
3	-26.5	11.7	I	-26.2	13.2	10.0	III
4	**-26.4**	**6.1**	**II**	**-26.1**	**2.5**	**8.1**	**I**
5	-26.3	12.4	I	-25.7	3.9	9.7	I
6	-26.3	6.7	II	-25.2	3.6	7.5	I
7	-26.2	6.7	II	-25.2	11.2	13.4	IV
8	-26.2	11.7	I	-24.8	10.5	10.3	II
9	-25.9	7.8	II	-24.8	3.0	6.8	I
10	**-25.9**	**6.3**	**II**	-24.7	13.3	8.2	III
11	-25.8	13.0	I	-24.7	3.0	4.1	I
12	-25.6	12.4	I	-24.3	10.5	13.1	II
13	-25.1	22.4	III	-24.1	9.5	12.3	II
14	-25.0	12.2	I	-23.6	9.1	9.5	II
15	-25.0	7.6	II	-23.5	9.5	9.9	II
16	-24.6	11.2	I	-23.4	9.7	12.1	II
17	-24.5	6.8	II	-23.1	11.0	14.1	II
18	-24.5	21.7	III	-22.8	11.5	11.8	IV
19	-24.4	11.7	I	-22.4	17.2	11.4	III
20	-24.2	11.8	I	-22.4	3.5	8.6	I

Table 5.3: Comparison of low energy nicotinamide and adenine dockings with the superimposed NAD-PT model. The twenty lowest energy clusters are shown for the adenine (total clusters: 94) and nicotinamide (total clusters: 99) dockings. The dockings used to construct the final NAD-PT models are shown in boldface. [a]See Table 5.2. [b]Root-mean-square distance between the heavy atoms of that docking and the corresponding atoms of the modeled NAD-PT complex (created by superposition of the NAD-DT docking reference - see Methods). [c]With few exceptions, the low energy dockings fall into one of three discrete regions of the PT surface, as shown in Figure 5.7. [d]See Table 5.3. Like DT, PT is photolabelled upon irradiation of the NAD-PT complex. However, for PT the reaction product has not been as well-characterized as that with DT. It seems likely that the resultant product is analogous to that of the NAD-DT reaction; thus we expect covalent bond formation between CG of decarboxylated Glu129 and C6 of the nicotinamide ring (C6N) of NAD. The Glu129:CG-NAD:C6N distance in the (superposition) NAD-PT model is 4.4Å.

8-azidoadenines seems much more likely for adenine derivatives bound to region I of DT (the nicotinamide subsite for NAD binding; Figure 5.6). Competitive binding studies with free nicotinamide and free adenine might clarify this issue. Second, it has long been known that affinity for DT follows the order adenine > adenosine > AMP ≈ ADP ≈ ATP (Kandel et al., 1974; Chung & Collier, 1977). The progressive *decrease* in affinity for DT that occurs for the series adenine→adenosine→AMP/ADP

(≈ 100:10:1) may be a manifestation of the ability of adenine to bind to both the nicotinamide and adenine subsites. This may not be possible for the larger fragments, due, perhaps, to steric or conformational problems.

5.3.3 Modeling the NAD-PT complex

The most exciting aspect of this work is the consistency of the adenine and nicotinamide fragment dockings to DT with the crystallographic NAD-DT structure, and the similarity (in essence, though not in detail) of the fragment dockings to DT and PT. Figure 5.7 shows the PT docking target with the sidechains of residues known to be crucial for NAD binding and/or catalytic activity highlighted, NAD from the superposition of the NAD-DT structure onto the PT target, and the low energy fragment dockings which again (similar to the results obtained with DT) cluster into discrete regions of the protein target surface. Glu129 of PT is thought to be the catalytic and structural equivalent of Glu148 of DT [evidence summarized in (Antoine et al., 1993)]. Superposition of the NAD-DT structure onto

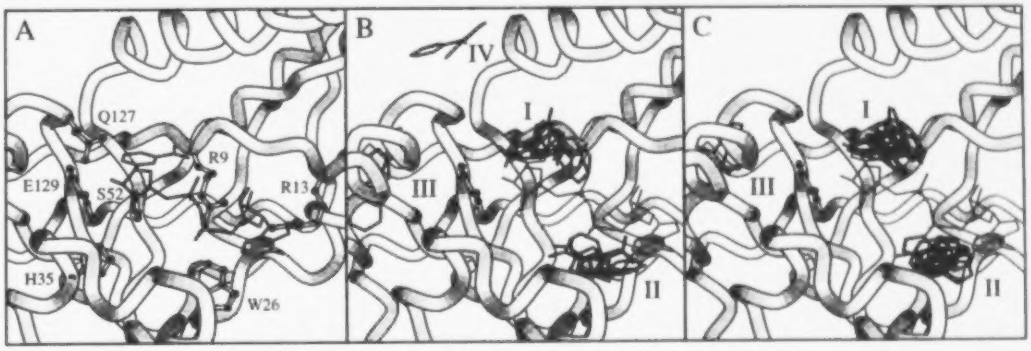


Figure 5.7: Low energy fragment dockings to PT. The modeled NAD binding site of PT is shown, with important sidechains (as for Figure 5.6 highlighted, and bound NAD from superposition of the NAD-DT structure (**A**). The low energy representatives of the twenty lowest energy docking clusters for nicotinamide (**B**) and adenine (**C**) are shown, along with the superimposed NAD molecule.

PT yields an NAD-PT complex in which the nicotinamide-PT interactions are similar to those of the NAD-DT complex, but the adenine-PT interaction differs significantly (Figure 5.7). Comparison of this model with the fragment docking results suggests that the modeled nicotinamide subsite is approximately correct, but that the adenine subsite of PT is shifted by approximately 5Å from that of DT (Table 5.3, Figure 5.7). The NAD-PT superposition model has one obvious clash between the Arg13 sidechain and the adenine moiety (Figure 5.7). However, this does not appear to be the reason for the shift (*versus* the dockings to DT) in the adenine fragment dockings, since the adenine dockings were essentially unaltered when an alternate rotamer, which avoided this clash, was chosen for this sidechain (results not shown).

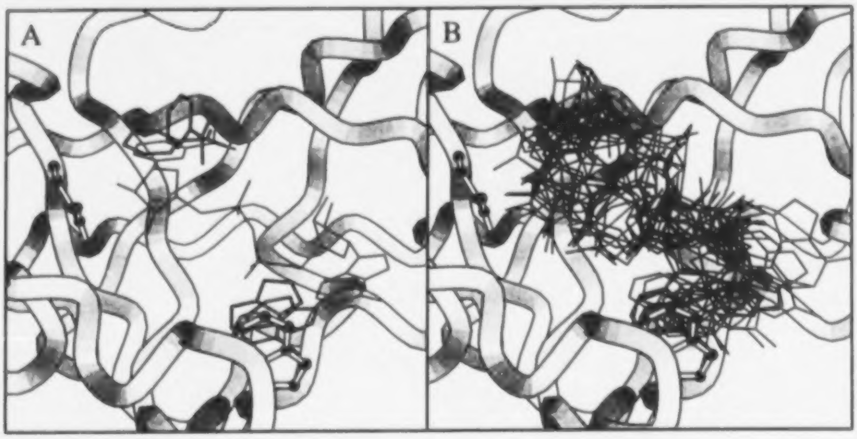


Figure 5.8: Modeling the NAD-PT complex. **A:** The low energy fragment dockings chosen for further modeling of the NAD-PT complex (see Table 5.3). NAD is from the superposition of the NAD-DT structure onto the PT docking target. **B:** The sixteen representative NAD dockings to PT, following Monte Carlo refinement (see Table 5.4).

Proceeding from the assumption that the adenine-PT fragment dockings represent the true adenine subsite of the NAD binding site of PT, and that the nicotinamide subsite of the NAD-PT superposition model is essentially correct, we visually examined the low energy fragment dockings to PT, and selected dockings that seem consistent with an NAD binding mode that utilizes these two subsites. Figure 5.8

shows the results of this selection process. We combined the two adenine and two nicotinamide fragments (Figure 5.8) to make the four possible adenine/nicotinamide pairs, and used five different topology files (described above) to generate different initial models of the NAD-PT complex. The twenty resultant models were further refined by flexible Monte Carlo docking/refinement. Twenty docking runs were performed for each of the twenty conformers, yielding a total of 389 dockings which passed the arbitrarily chosen interaction energy cutoff (-20.0 kcal mol^{-1}). A single output list was created containing all 389 NAD-PT models; cluster analysis with distance criteria of 2 and 1.5Å RMS gave 11 and 16 clusters, respectively. We used the more stringent criterion, and visually inspected the 16 representative dockings (Figure 5.8).

docking	energy[a] (kcal mol^{-1})	photolabeling distance[b] (Å)
1	-69.8	4.9
2	-68.8	4.7
3	-63.6	8.8
4	-58.7	6.1
5	-56.3	6.6
6	-55.7	6.5
7	-54.3	9.9
8	-52.7	9.2
9	-49.6	4.0
10	-46.5	4.8
11	-44.6	4.7
12	-41.2	5.1
13	-37.2	8.0
14	-36.7	4.3
15	-26.3	7.1
16	-23.3	9.7

Table 5.4: Refined models of the NAD-PT complex. [a]The models were generated by flexible superposition, rigid-body conjugate gradient minimization, and flexible Monte Carlo docking/refinement, as described in the text. The interaction energies shown were calculated after Monte Carlo refinement. [b]See Table 5.3.

Like the photocrosslinking reaction that occurs between NAD and Glu148 of DT (see above; (Carroll et al., 1985)), Glu129 of PT is photolabelled by the nicotinamide moiety of NAD (Cockle, 1989). Although the reaction product has not been as well-characterized as that of the NAD-DT reaction, it seems likely that the analogous reaction occurs. We therefore expect decarboxylation of Glu129, and covalent bond formation between Glu129:CG and C6N of the nicotinamide ring. The Glu148:CG-C6N distance in the NAD-DT complex is 4.1Å, and the analogous Glu129:CG-C6N distance in the NAD-PT (superposition) model is 4.4Å. This distance can be used as a crude distance filter to gauge the plausibility of a given model of the NAD-PT complex. On the basis of this distance, and the calculated interaction energies, dockings 1 and 2 seem the most reasonable of the lower energy models (Table 5.4).

docking 1		docking 2	
PT residue	energy[a] (kcal mol^{-1})	PT residue	energy (kcal mol^{-1})
Arg:9	-18.0	Arg:9	-13.7
Tyr:10	-4.0	Tyr:10	-4.9
Arg:13	-8.2	Arg:13	-7.4
Thr:24	-2.1	Trp:26	-10.0
Trp:26	-9.4	Val:51	-2.4
His:35	-1.7	Ser:52	-4.2
Ser:52	-3.0	Thr:53	-2.2
Tyr:59	-2.2	Tyr:59	-2.8
Arg:67	-1.7	Arg:67	-2.8
Gln:127	-3.6	Gln:127	-2.9
total	-65.1	total	-63.2

Table 5.5: Important intermolecular contacts for dockings 1 and 2. Three rounds of progressively less restrained energy minimization were performed, with the DISCOVER implementation of the AMBER potential energy function (see Methods) [a]The calculated interaction energies for the ten most important NAD-PT-residue contacts, as well as the total interaction energies for each docking.

Table 5.5 shows the more significant interactions between NAD and PT for dockings 1 (see also Table 5.6) and 2, on a residue-by-residue basis. Bell & Eisenberg (1996) made some general predictions about the residues involved in NAD binding to PT, based on sequence alignment and the NAD-DT structure; our more specific

NAD-PT models (dockings 1 and 2) are, in general, consistent with their predictions. Since catalysis involves cleavage of the *N*-glycosidic bond (this is related to the photolabeling reaction discussed above), the positioning of the nicotinamide moiety must be especially critical. Such reasoning is supported by the remarkably similar orientation of this moiety in the hNAD-ETA and NAD-DT structures (Figures 5.4 and 5.5). This feature is conserved in our NAD-PT models with dockings 1 and 2 (Figure 5.9, Table 5.6). Hydrogen bonds between PT and NAD:O7N and NAD:N7N in both models are identical to those observed in the NAD-DT and hNAD-ETA structures. The importance of these two interactions was first suggested by the

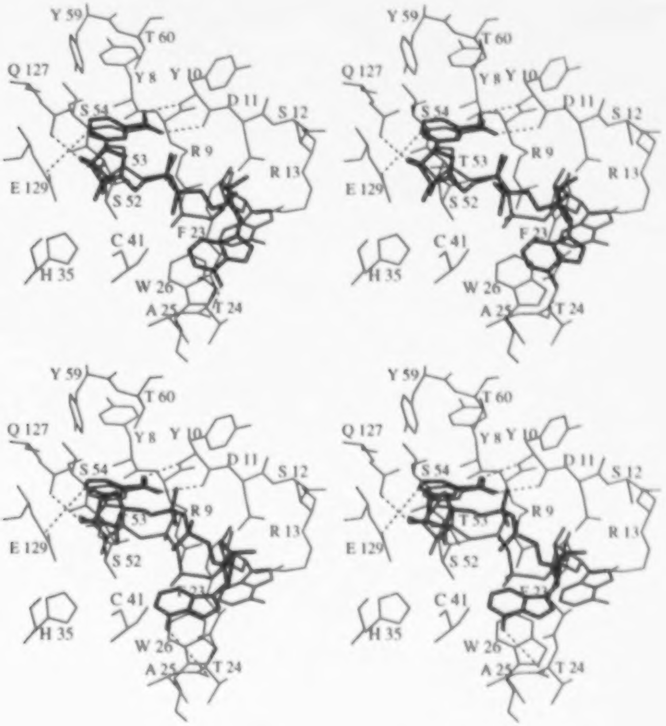


Figure 5.9: Two low energy models of the NAD-PT complex. The complexes involving NAD dockings 1 (upper) and 2 (lower), showing most of the residues with one or more atoms within 8Å of the ligand. For docking 1, possible intermolecular hydrogen bonds are listed in Table 5.6.

results of competition studies with DT and NAD analogs; these results showed that slight modifications ($-CONH_2 \rightarrow -CHO$, $-COCH_3$) of the exocyclic amide moiety had marked effects on NAD binding [(Kandel et al., 1974; Lory et al., 1980), but see also (Kessler & Galloway, 1992)]. The positioning of the nicotinamide-ribose fragment relative to Glu129 is very similar to that observed for Glu148 and Glu553 in the NAD-DT and hNAD-ETA structures, respectively. The conformation of the nicotinamide-ribose (O4Y-C1Y-N1N-C2N) linkage is eclipsed in the NAD-DT structure; it is thought that this strain could contribute to the lability of this bond, and thus may be an important feature of catalysis (Bell & Eisenberg, 1996). Docking 1 has a similarly eclipsed O4Y-C1Y-N1N-C2N dihedral; docking 2 does not, but several of the other lower energy dockings do [note that this dihedral was free to rotate in some of the flexible superpositions (Table 5.1), and in all of the flexible Monte Carlo refinements].

donor	acceptor	D···A[a] distance (Å)	D···H···A angle (degrees)
Arg9:NH1	NAD:O1N	2.7	135
Arg9:NH2	NAD:O1N	2.7	132
Tyr10:N	NAD:O7N	2.8	169
NAD:N7N	Tyr10:O	3.2	173
NAD:O2X	Asp11:OD2	3.1	109
Arg13:NH1	NAD:O2X	2.9	137
Arg13:NH2	NAD:O2X	2.1	134
NAD:N6A	Thr24:O	2.9	147
Ser52:OG	NAD:N1N	3.1	147
Gln127:NE2	NAD:O2Y	3.1	148

Table 5.6: Important intermolecular hydrogen bonds for docking 1. [a]D: donor, A: acceptor; we used a relatively permissive hydrogen bond filter of 90° for the D···H···A angle, and 3.0 Å for the H···A distance.

In contrast to the nicotinamide-ribose portion of the NAD-PT models, the adenosine diphosphate moiety of these models is in a significantly different conformation from that of both the NAD-DT complex and our model of the

NAD-ETA complex. In the experimental structure two direct NAD-DT hydrogen bonds, DT:Gln36:N-NAD:N1A and DT:Gly34:O-NAD:N6A, stabilize the orientation of the adenine moiety of bound NAD. Similar interactions are observed in the hNAD-ETA structure [(Li et al., 1995); Figure 5.5], as well as in our model of the NAD-ETA complex (Figure 5.5). The NAD-PT complex with docking 1 has a similar PT:Thr24:O-NAD:N6A hydrogen bond, but the DT:Gln36:N-NAD:N1A interaction is not conserved. Docking 1 is more compact (16.9Å in length) than the NAD conformation in the NAD-DT structure (19.9Å in length). One important consequence of these differences is that the solvent-exposed NAD surface in the NAD-PT model is quite different from that of the NAD-DT (and NAD-ETA) structures; however, in all three complexes the ribose and phosphate moieties are largely solvent exposed [(Bell & Eisenberg, 1996); Figures 5.4, 5.5, and 5.9]. Collier and co-workers (Kandel et al., 1974) first noted that DT-catalyzed ADP-ribosylation decreased with increasing ionic strength, whereas NAD hydrolysis (and therefore, by inference, NAD binding to DT) was not affected [a comparatively minor effect was reported recently for the PT-catalyzed ADP-ribosylation of a synthetic peptide consisting of the C-terminal twenty residues of $G_{i\alpha 3}$ (Finck-Barbancon & Barbieri, 1995)]. This salt effect is entirely consistent with the experimental NAD-DT structure (Bell & Eisenberg, 1996), and our modeled NAD-ETA structure (see above), in which the ribose and phosphate moieties are largely solvent-exposed. Most of the crucial intermolecular interactions of these complexes involve the adenine or nicotinamide moieties of the ligands, and not the highly polar/charged ribose and phosphate groups. On the other hand, as was recently proposed (Bell & Eisenberg, 1996), this region of the NAD-DT surface may be essential to the recognition of the NAD-DT complex by EF-2. EF-2 will not bind *apo*-DT; ADP-ribosylation follows an ordered sequential mechanism involving an initial NAD-DT complex, followed by formation of the NAD-DT-EF-2 ternary complex (Chung & Collier, 1977). The importance of charge-charge interactions to the latter recognition event is supported by the salt effects described above. Furthermore, the target Cys residue of the $G_{\alpha i}$ substrate of PT is within a few residues of two positively charged sidechains (. . .KNNLKE**C**. . .), and the diphthamide sidechain target of DT and ETA has a positively charged tertiary amine that would be within 10Å of the nucleophilic N that is ADP-ribosylated. Interactions between these positively charged moieties near the ADP-ribosylation targets and negatively charged groups of the NAD-toxin complexes (*eg.* the diphosphate group) may be important in stabilization of the ternary complex. These speculations are also consistent with the fact that both DT and ETA have the same ADP-ribosylation target protein, EF-2, whereas PT recognizes different target substrates (G proteins). Differences in the relevant region of the NAD-toxin surface, due in part to differences in NAD conformation, may contribute to discrimination between different ADP-ribosylation protein targets. Thus, a significant difference between the conformation of NAD bound to PT and that of NAD bound to DT or ETA may be essential to function.

5.3.4 A note on ligand design

Along with the known structural and biochemical information that provided the foundation for our modeling studies, these results and analyses have direct bearing on the design of small molecule inhibitors of PT, as well as other related ADPR toxins. The ability of adenine to bind to two subsites of the NAD binding site suggests a di-adenine dinucleotide as one reasonable starting point. An adenine analog modified to mimic the obviously important interactions that occur between the exocyclic amide of the nicotinamide ring and the toxins (in our NAD-PT model as well as in the NAD-DT and hNAD-ETA structures) would seem to be another. We are unaware of any reports of the affinity of NADH for PT, but the reduced dinucleotide competes with NAD for DT binding with high affinity (Goor & Maxwell, 1970). Visual inspection of the solvent-accessible surfaces of the nicotinamide subsites of the complexes shows a very tight-fitting pocket in all cases (not shown). If these sites can accomodate the presumably non-planar nicotinamide ring of NADH as well as that of NAD, a variety of (modified) ring systems may be useful as starting points for analog design. A small nicotinamide analog that mimics the interactions of the exocyclic amide group in the known and modeled structures, and that also interacts with Gln127 and Ser52, as suggested by our model (Table 5.5, Figure 5.9), might be the target molecule to focus on. The similar affinity of NAD and NADH suggests that modifications at C4N of the nicotinamide ring might be also be fruitful, although the well-defined and relatively restricted binding pockets would seem to argue against this.

5.3.5 Significance

We have proposed a model of the NAD-PT complex, based on the structures of the related NAD-DT and hNAD-ETA complexes. There are important similarities and differences between the way in which NAD binds to PT in our model, and to both DT and ETA in the known structures. This structural difference may be important for proper function, particularly in substrate recognition. Also, we have shown that our model is consistent with a variety of experimental results.

Construction of our model(s) of the NAD-PT complex, as well as the other studies presented here, involved a novel application of the fragment-based docking approach to ligand design. Our fragment docking results were quite striking, in their obvious relation to the known NAD-DT and hNAD-ETA structures. These results offer a clear indication of the usefulness of these particular computational tools, and the approach we have described, in solving problems of structure prediction. This is in agreement with the increasing frequency of reports that describe the usefulness of docking tools in a variety of related applications.

The ADPR toxins are a medically important class of enzymes; detailed understanding of their structure, function, and mechanisms of action are therefore vitally important. One potential use of the structural information central to the present work is in the design of small molecule competitive inhibitors of NAD binding. Our docking study has involved the analysis and discussion of a variety of biochemical and structural information directly relevant to this challenging problem, and we have summarized a few molecular principles that may be useful in this particular design process.

5.4 References

Antoine, R., Tallet, A., van Heyningen, S., & Locht, C. (1993). Evidence for a catalytic role of glutamic acid 129 in the NAD-glycohydrolase activity of the pertussis toxin subunit. *J. Biol. Chem.*, 268:24149–24155.

Bass, M. B., Hopkins, D. F., Jaquysh, W. A. N., & Ornstein, R. L. (1992). A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. *Proteins: Structure, Function, and Genetics*, 12:266–277.

Bell, C. E. & Eisenberg, D. (1996). Crystal structure of diphtheria toxin bound to nicotinamide adenine dinucleotide. *Biochemistry*, 35:1137–1149.

Bennett, M. J., Choe, S., & Eisenberg, D. (1994). Refined structure of dimeric diphtheria toxin at 2.0 Å resolution. *Protein Sci.*, 3:1444–1463.

Bennett, M. J. & Eisenberg, D. (1994). Refined structure of monomeric diphtheria toxin at 2.3 Å resolution. *Protein Sci.*, 3:1464–1475.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542.

Blanke, S. R., Huang, K., & Collier, R. J. (1994a). Active-site mutations of diphtheria toxin: role of tyrosine-65 in NAD binding and ADP-ribosylation. *Biochemistry*, 33:15494–15500.

Blanke, S. R., Huang, K., Wilson, B. A., Papini, E., Covacci, A., & Collier, R. J. (1994b). Active-site mutations of diphtheria toxin catalytic domain: role of histidine-21 in nicotinamide adenine dinucleotide binding and ADP-ribosylation of elongation factor 2. *Biochemistry*, 33:5155–5161.

Bohm, H.-J. (1992). The computer program ludi: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Design*, 6:61–78.

Brandhuber, B. J., Allured, V. S., Falbel, T. G., & McKay, D. B. (1988). Mapping the enzymatic active site of *Pseudomonas aeruginosa* exotoxin A. *Proteins: Struct. Funct. Genet.*, 3:146–154.

Carroll, S. F., McCloskey, J. A., Crain, P. F., Oppenheimer, N. J., Marschner, T. M., & Collier, R. J. (1985). Photoaffinity labeling of diphtheria toxin fragment A with NAD: structure of the photoproduct at position 148. *Proc. Natl. Acad. Sci. USA*, 82:7237–7241.

Chung, D. W. & Collier, R. J. (1977). The mechanism of ADP-ribosylation of elongation factor 2 catalyzed by fragment A from diphtheria toxin. *Biochim. Biophys. Acta*, 483:248–257.

Cockle, S. A. (1989). Identification of an active-site residue in subunit S1 of pertussis toxin by photocrosslinking to NAD. *FEBS Lett.*, 249:329–332.

Cortina, G. & Barbieri, J. T. (1991). Localization of a region of the S1 subunit of pertussis toxin required for efficient ADP-ribosyltransferase activity. *J. Biol. Chem.*, 266:3022–3030.

Cummings, M. D., Hart, T. N., & Read, R. J. (1995). Monte Carlo docking with ubiquitin. *Protein Sci.*, 4:885–899.

DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D., & Venkataraghavan, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem*, 29:2149–2153.

Domenighini, M., Magagnoli, C., Pizza, M., & Rappuoli, R. (1994). Common features of the NAD binding and catalytic site of ADP-ribosylating toxins. *Mol. Microbiol.*, 14:41–50.

Domenighini, M., Montecucco, C., Ripka, W. C., & Rappuoli, R. (1991). Computer modelling of the NAD binding site of ADP-ribosylating toxins: active-site structure and mechanism of NAD binding. *Mol. Microbiol.*, 5:23–31.

Finck-Barbancon, V. & Barbieri, J. T. (1995). Adp-ribosylation of α_{i3}c20 by the S1 subunit and deletion peptides of S1 of pertussis toxin. *Biochemistry*, 34:1070–1075.

Goor, R. S. & Maxwell, E. S. (1970). The diphtheria toxin-dependent adenosine diphosphate ribosylation of rat liver aminoacyl transferase II. *J. Biol. Chem.*, 245:616–623.

Hart, T. N. & Read, R. J. (1992). A multiple-start Monte Carlo docking method. *Proteins: Structure, Function, and Genetics*, 13:206–222.

Hart, T. N. & Read, R. J. (1994). Multiple-start Monte Carlo docking of flexible ligands. In Merz, K.M., Jr. & LeGrand, S., editors, *The Protein Folding Problem and Tertiary Structure Prediction*, pages 71–108. Birkhäuser, Boston.

Johnson, V. G. & Nicholls, P. J. (1994). Histidine 21 does not play a major role in diphtheria toxin catalysis. *J. Biol. Chem.*, 269:4349–4354.

Kandel, J., Collier, R. J., & Chung, D. W. (1974). Interaction of fragment A from diphtheria toxin with nicotinamide adenine dinucleotide. *J. Biol. Chem.*, 249:2088–2097.

Kessler, S. P. & Galloway, D. R. (1992). *Pseudomonas aeruginosa* exotoxin A interaction with eucaryotic elongation factor 2. Role of the His^{426} residue. *J. Biol. Chem.*, 267:19107–19111.

Li, M., Dyda, F., Benhar, I., Pastan, I., & Davies, D. R. (1995). The crystal structure of *Pseudomonas aeruginosa* exotoxin domain III with nicotinamide and AMP: Conformational differences with the intact exotoxin. *Proc. Natl. Acad. Sci. USA*, 92:9308–9312.

Lory, S., Carroll, S. F., Bernard, P. D., & Collier, R. J. (1980). Ligand interactions of diphtheria toxin I. Binding and hydrolysis of NAD^+. *J. Biol. Chem.*, 255:12011–12015.

Merritt, E. A. & Hol, W. G. J. (1995). AB_5 toxins. *Curr. Op. Struc. Biol.*, 5:165–171.

Michel, A. & Dirkx, J. (1977). Occurrence of tryptophan in the enzymatically active site of diphtheria toxin fragment A. *Biochim. Biophys. Acta*, 491:286–295.

Moon, J. B. & Howe, W. J. (1991). Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *Proteins: Struct, Funct, Genet*, 11:314–328.

Moss, J., Garrison, S., Oppenheimer, N. J., & Richardson, S. H. (1979). NAD-dependent ADP-ribosylation of arginine and proteins by *Escherichia coli* heat-labile enterotoxin. *J. Biol. Chem.*, 254:6270–6272.

Moss, J. & Vaughan, M. (1988). ADP-ribosylation of guanyl nucleotide-binding regulatory proteins by bacterial toxins. *Adv. Enzymol.*, 61:303–379.

Oppenheimer, N. J. (1978). Structural determination and stereospecificity of the choleragen catalyzed reaction of NAD+ with guanidines. *J. Biol. Chem.*, 253:4907–4910.

Papini, E., Santucci, A., Schiavo, G., Domenighini, M., Neri, P., Rappuoli, R., & Montecucco, C. (1991). Tyrosine 65 is photolabeled by 8-azidoadenine and 8-azidoadenosine at the NAD binding site of diphtheria toxin. *J. Biol. Chem.*, 266:2494–2498.

Read, R. J. & Stein, P. E. (1993). Toxins. *Curr. Op. Struc. Biol.*, 3:853–860.

Rotstein, S. H. & Murcko, M. A. (1993). Groupbuild: A fragment-based method for *de novo* drug design. *J. Med. Chem.*, 36:1700–1710.

Soman, G., Narayanan, J., Martin, B. L., & Graves, D. J. (1986). Use of substituted (benzylidineamino)guanidines in the study of guanidino group specific ADP-ribosyltransferase. *Biochemistry*, 25:4113–4119.

Stein, P. E., Boodhoo, A., Armstrong, G. D., Cockle, S. A., Klein, M. H., & Read, R. J. (1994). The crystal structure of pertussis toxin. *Structure*, 2:45–57.

Stoddard, B. L. & Koshland, D. E. J. (1992). Prediction of the structure of a receptor-protein complex using a binary docking method. *Nature*, 358:774–776.

Weiner, S., Kollman, P., Case, D., Singh, U., Ghio, C., Alagona, G., Profeta, S. J., & Weiner, P. (1984). A new force-field program for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784.

Wilson, B. A., Blanke, S. R., Reich, K. A., & Collier, R. J. (1994). Active-site mutations of diphtheria toxin: tryptophan 50 is a major determinant of NAD affinity. *J. Biol. Chem.*, 269:23296–23301.

Wilson, B. A., Reich, K. A., Weinstein, B. R., & Collier, R. J. (1990). Active-site mutations of diphtheria toxin: effects of replacing glutamic acid-148 with aspartic acid, glutamine, or serine. *Biochemistry*, 29:8643–8651.

Chapter 6

General Discussion and Conclusions

Each chapter of this dissertation is relatively self-contained, and the conclusions specific to each have been stated in the respective Discussion and/or Conclusion sections; Chapters 2 and 3, especially, contain lengthy discussions and summaries. More specifically, the *biochemical problems* of interest are each confined to separate chapters, with the exception of the diubiquitin system which was studied in detail in Chapter 2, and was used as one of the test systems in Chapter 3. However, all of the problems investigated involve molecular recognition, and it is the methods used in studying these processes that relate the various chapters of this thesis. As stated earlier, much of the work described here can be described as "methods testing"; this statement applies particularly to Chapters 2 and 3. For the most part, therefore, this final section will be confined to a discussion of some of the general issues involved in automated docking and molecular modeling.

The results of a protein-protein "docking challenge" that was posed to the community of researchers in the field were recently published (Strynadka et al., 1996), and they are quite encouraging. Five of the six groups that took up the challenge produced "best" answers that fit into the category of "near to correct" which we outlined in Chapter 3 (docking probe between 2 and 5Å RMS away from the correct answer). In this study answers ranging from 3.35Å to 6.11Å RMS from the crystallographic structure were considered to be correct; this suggests that our criterion (outlined in Chapters 2 and 3) for correctness was, perhaps, too strict. For example, dockings 1 and 3 with two copies of monoubiquitin (Table 2.6) are clearly correct according to this relaxed criterion, and this revised conclusion is consistent with the opinions of several authorities in the field (Strynadka et al., 1996). It seems reasonable to state that protein-protein docking, involving a relatively large docking probe and probe-target interface, requires a more relaxed definition of correctness than that required for protein-small molecule docking. In the latter case a difference of a few Å RMS can describe rotations of 180° and completely different atom-atom contacts; this is not true for protein-protein docking. In the docking challenge (Strynadka et al., 1996), none of the entries, which involved both rigid and flexible docking methods and a variety of different scoring schemes (Strynadka et al., 1996), gave good predictions of the precise atom-atom interactions at the active site. Consistent with our results involving sidechain truncation, and similar results reported by other groups with other docking systems (discussed in Chapters 1- 3 and Appendix B), Strynadka et al. (1996) conclude that gross matching of surfaces, rather than accurate modeling of intermolecular atom-atom contacts, is sufficient for reasonably accurate predictions of protein-protein associations.

Ideally, an automated docking simulation would search all possible configurations of the complex of interest and pick the correct one to be the one of lowest energy. Furthermore, this conclusion would be arrived at without considering any additional information (*eg.* binding or mutation studies, chemical modifications, mechanistic details). Current methods do not allow for this ideal experiment due to a variety of limitations.

Drug design is one of the common applications of docking simulations. Consideration of a typical drug design scenario, however, leads one to the conclusion that such a powerful method is not strictly necessary (although it is, of course, desirable). Simply speaking, in this scenario the investigators will have a target structure derived from either experiment or calculation, several structurally-related ligands which exhibit a wide range of affinities for the target, and some information regarding the nature of the site of interaction (from, for example, competition, mutation, or chemical modification studies). This type of information can be incorporated into a docking study to greatly reduce the configurational (and/or conformational) space which must be searched during the docking simulation. This in turn will allow for a more exhaustive search of the limited space, thereby increasing the chances of determining the correct binding mode(s). Alternatively, such information can be applied as a filter to reject some of the data obtained in an unrestricted docking simulation. *A significant aspect of the work described in the preceding chapters has been the application of non-energetic information to the analysis of the results of energy-based docking simulations.* This statement applies particularly to Chapters 2 and 3. We have also found such information to be helpful in the manual modeling studies described in Chapters 4 and 5 (Chapter 5 involved both automated docking and manual modeling).

In our docking simulations, dockings were ranked according to the interaction energies calculated with a relatively simple potential function (discussed in Chapters 1 and 3 and Appendix A). For systems with known answers, where we were testing some aspect of our docking procedure (Chapters 2, 3, and, to some extent, 5), our first step was generally to evaluate our ability to achieve the ideal ranking described above (paragraph 3 of this section). In some cases this ideal ranking was clearly achieved, with the correct docking having a substantially lower interaction energy than any of the incorrect dockings. In other cases, in our own work as well as in reports from many other groups (citations in Chapters 1- 3 and Appendix B) energy-based ranking does not clearly discriminate between correct and incorrect dockings. A variety of limitations, both practical and theoretical, are responsible for this failing. However, it is often possible to overcome the limitations of current methods. Non-energetic information relevant to this discrimination is often available, and its application to the problem can be straightforward and simple [*eg.* (Strynadka et al., 1996), and the preceding chapters]. We have shown that in some cases the resulting clarification can be dramatic. Several methods that we have found useful are summarized below.

- **Sidechain truncation.** In Chapter 2 we showed that truncation of a flexible Arg sidechain allowed for prediction of diubiquitin using two copies of (otherwise) native monoubiquitin. Information was available, from both structural and chemical modification studies, that indicated that this (Arg42) and several other sidechains might be crucial to the diubiquitin interface. Truncation is a crude and extreme approximation of sidechain flexibility; in Chapter 2 and Appendix B more realistic alternatives are discussed. However, if protein-protein association can be reasonably approximated by relatively gross surface matching techniques, limited sidechain truncation may remain a useful approximation for some time to come.

- **Distance constraints.** Information regarding specific intermolecular atom-atom or residue-residue contacts involved in the complex of interest is often available. This may be derived from a variety of experimental techniques (*eg.* chemical modification, site-specific mutagenesis, binding studies). Such information can be *extremely* useful in computational docking studies, for both experimental design and analysis of results. For example, in Chapter 2 we used the requirement of covalent bond formation between the two halves of diubiquitin to rule out an energetically favorable but incorrect cluster of dockings, and in Chapter 4 we used the observed changes in binding affinity for a doubly-mutated lectin to choose between several possible ligand binding sites, and then built a model of the wild-type protein bound to its preferred ligand.

- **Nature of the buried surface.** The ranking of dockings with monoubiquitin was dramatically improved by correcting the final calculated interaction energies with an ASP-based desolvation term. Upon further investigation of this procedure we found that it was difficult to choose between the many different reported procedures. This led us to re-determine several parameter sets, and then empirically evaluate the usefulness of the ASP sets in the analysis of several different protein-protein docking studies. The results presented in Chapter 3 clearly show that the ASP set derived from octanol-water transfer energies is most appropriate for our procedure. A recent study involving a similar comparison arrived at the same conclusion (Juffer et al., 1995).

Given recent progress, it seems reasonable to expect that some of these limitations will be overcome in the next few years due to increases in available computer power, and/or implementation of algorithms or methods which account for molecular flexibility and desolvation. These methods are already appearing (discussed and cited in Appendix B). Some of the approximations that are currently necessary and useful will no longer be required with the more sophisticated procedures.

Methods incorporating flexibility for all or selected sidechains have been appearing with increasing frequency in the literature (Appendix B), although flexibility is often limited to the docking probe. Methods that include backbone flexibility are only beginning to emerge (Abagyan & Totrov, 1994; Totrov & Abagyan, 1994); to date these methods show great promise, and it will be interesting to see them tested with native proteins that undergo major conformational adjustment(s) upon binding. In our study with monoubiquitin we showed that the conformation of one Arg sidechain prevented generation of a diubiquitin-like dimer with rigid-body docking (Chapter 2). The sidechain rotamers observed in both the mono- and diubiquitin structures were present in a very limited rotamer library (results not shown), suggesting that a very simple rotamer-based sidechain search would suffice, at least for this system. Docking two copies of native monoubiquitin to generate a diubiquitin-like structure should provide a good test case for docking methods that allow sidechain flexibility in both the probe and target molecules.

Consideration of explicit bulk solvent in automated docking procedures will undoubtedly remain impractical for some time. It is not yet clear what the most appropriate method for *approximating* bulk solvent effects is; however, a general discussion of this challenging issue is well beyond the scope of this dissertation. We studied one method of approximating bulk desolvation, and its applicability to our docking method. Several different parameter sets were tested with the surface-area-based method developed by Eisenberg & Mclachlan (1986). In our procedure, wherein a desolvation term was added (only for the final energy calculation) to a standard Lennard-Jones plus Coulomb potential function, van der Waals interactions are "double-counted" for the ASP sets derived from solution-solution transfer data. However, the ASP set derived from hydration (vacuum-water transfer) energies significantly disturbed the rankings we tested, whereas the parameter set derived from octanol-water transfer seemed consistent with our energy calculation. A simple energy function that considers intermolecular van der Waals and hydrogen-bonding interactions, unsatisfied burial of hydrogen-bonding groups, and desolvation of hydrophobic surface might be useful for automated docking. Recent reports show that simplified molecular representations and scoring schemes (Vakser & Aflalo, 1994; Gehlaar et al., 1995) can be as successful as much more sophisticated methods in predicting the structures of complexes. However, application of automated docking to many aspects of drug design requires a reasonable degree of accuracy in the *relative* ranking of ligands; this must at least be true for structurally related ligands, and accuracy over a range of structurally diverse molecules is even more desirable.

The preceding discussion in this section has focused on the application of automated molecular docking to the *ab initio* prediction of bimolecular complexes - the ultimate goal of development in this field. However, docking and other molecular modeling techniques can be useful in other situations as well. The studies described in Chapters 4 and 5 involved the application of rigid and flexible superposition, manual model building, systematic conformational search, and automated docking to protein-small molecule structure prediction. In conjunction with relevant structural and biochemical information, these computational tools allowed us to make reasonable predictions of the structures of two complexes.

A general discussion of the "principles" of manual modeling does not seem possible, since significant components of the process are personal intuition and bias. Certainly, many of the principles discussed above apply. In our work (Chapters 4 and 5) we used various automated or semi-automated procedures to construct a relatively small number of initial models. Many of the analysis tools developed for automated docking were useful in discriminating between good and bad initial models. When we were able to focus on one or a few possible models, the models were adjusted manually, and further refined with constrained/restrained energy minimization.

A rigid-spheres method was used to generate several likely conformations of the tri- and tetrasaccharide moieties of the cell-surface glycolipid receptors for native and mutant lectin subunits of pig edema toxin (Chapter 4). Examination of published binding data obtained with a variety of mutants allowed us to make an educated guess as to which of three observed binding sites for a related protein was most likely to be relevant to the double mutant of interest. We then superimposed the related binding site of interest onto the mutant pig edema toxin, modeled the wild type protein, and constructed a model of the lectin-carbohydrate complex that offers a reasonable explanation of the difference in binding preferences between the native and mutant toxin. Our model of the complex is consistent with the results of many binding studies, and also suggests a major role for one residue which has not yet been studied by site-specific mutagenesis.

In this study, the binding data obtained with different mutants was very useful in directing our attention to one of the three modeled (from observed sites for a related protein) Gb3/Gb4 binding sites. Inspection of the three sites with several important sidechains highlighted (Figure 4.4) made the choice simple; since the biological relevance of the different observed sites is not clear, this observation made our task significantly easier. The final results of this modeling study support this initial choice. However, this raises an important general issue regarding the interpretation of binding experiments in terms of distance constraints for docking simulations. When site-specific mutations have marked effects on binding, or when binding affects some residue-based phenomenon (*eg.* fluorescence), the simplest interpretation is commonly invoked: the residue in question is at or near the binding site. This interpretation can be misleading, and the importance of any one residue or ligand moiety should only be inferred in the context of all the available and relevant information. Our study of Gb3/Gb4 binding shows that information for several residues can be quite conclusive; however, some of the residues were within similar distances of all three modeled binding sites. When considered alone, these residues would not have been helpful in discriminating between the three sites. Similarly, in our study with diubiquitin, crucial residues had been identified near the dimer interface, as well as relatively distant from it. In this case the known variant residues were also informative, in that these residues were all distant from the dimer interface. Finally, this type of confusion also applied, to some extent, to our study of the NAD-PT complex (see below and Chapter 5).

Oligosaccharides represent a particular challenge for molecular modeling. The inter-residue conformational preferences are not nearly as sharply defined as for peptides, and the energy barriers are much smaller. The relationship between the conformational preferences of free and bound oligosaccharides is also unclear. These difficulties are compounded by the relative abundance of mobile polar hydrogens in these molecules. Our manual modeling study with Gb3 and Gb4 suggests that simple methods for predicting saccharide linkage conformation may be inadequate for automated docking of these molecules. A reasonable approach to this problem might be to select two or three likely regions of ϕ-ψ space for a given linkage, and then perform some type of focused search around these regions during the docking simulation. For the hydroxyl groups, a relatively coarse initial search should suffice, perhaps followed by optimization of the best position thus obtained. Alternatively, this latter problem has been addressed in a different way. In the CHEAT (Carbohydrate Hydroxyl groups represented by Extended ATоms) approach (Grootenhuis & Haasnoot, 1993), carbohydrate hydroxyls are represented as united (extended) atoms, much like the aliphatic carbons in many of the more popular potential functions. Perhaps a similar method may be developed for other types of mobile polar hydrogen atoms.

In our final study we modeled the complex of NAD bound to the ADP-ribosyltransferase subunit of pertussis toxin (PT). This was based on two related structures with diphtheria toxin (DT), and one with *Pseudomonas* exotoxin A (ETA). In generating NAD conformers for rigid-body docking, we again encountered the difficulty posed by even relatively small carbohydrate moieties. We used information from the available complex structures, as well as new flexible superposition and docking tools (unpublished programs of Trevor Hart), to surmount this difficulty.

During rigid-body docking studies with various sizes of NAD fragments, we observed marked clustering of the low energy dockings of nicotinamide and adenine. For DT these clusters were related to the NAD-DT structure in an obvious way (Figure 5.4). Most of the intermolecular contacts in the NAD-DT complex involve the adenine or nicotinamide moieties of NAD, with the ribose and phosphate groups being largely solvent-exposed (Bell & Eisenberg, 1996). Furthermore, both adenine and nicotinamide are relatively rigid, and for our purposes one conformer for each fragment provides adequate representation. Therefore, by "building up" the conformation(s) of NAD bound to PT (or DT or ETA) we could avoid many of the difficulties associated with the conformational complexity of the NAD molecule, and yet not compromise our final model with this approximation. During the course of this study a new set of flexible tools became available to us, and were particularly useful in this NAD "build-up" procedure.

These results are particularly relevant to the prediction of the structures of biomolecular complexes. Several recent reports (discussed in Chapters 1 and 3) have indicated that interactions between hydrophobic groups are of particular importance in these associations. Our fragment-based docking results with NAD and PT, as well as the related complex structures, are consistent with this conclusion (of course hydrophobicity is a relative scale, but nicotinamide and adenine appear to be less polar than the remainder of the NAD molecule). Just as gross surface recognition may be sufficient for approximately correct protein-protein recognition (see above), hydrophobic surfaces may be of primary importance in the recognition of smaller ligands.

Fragment-based docking methods are often used as the foundation for fragment-library-based "build-up" methods for *ab initio* ligand construction. Our results with NAD docking to PT, DT, and ETA suggest that hydrophobic fragments may be the most efficient starting point for such methods of ligand construction. Of course, our results were obtained with a few closely related systems; a much broader survey is required to determine the general applicability of this principle.

6.1 References

Abagyan, R. & Totrov, M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, 235:983–1002.

Bell, C. E. & Eisenberg, D. (1996). Crystal structure of diphtheria toxin bound to nicotinamide adenine dinucleotide. *Biochemistry*, 35:1137–1149.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, 319:199–203.

Gehlaar, D. K., Verkhivker, G. M., Retjo, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J., & Freer, S. T. (1995). Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry & Biology*, 2:317–324.

Grootenhuis, P. D. J. & Haasnoot, C. A. G. (1993). A charmm based force field for carbohydrates using the cheat approach: carbohydrate hydroxyl groups represented by extended atoms. *Molec. Simulat.*, 10:75–95.

Juffer, A. H., Eisenhaber, F., Hubbard, S. J., Walther, D., & Argos, P. (1995). Comparison of atomic solvation parametric sets: Applicability and limitations in protein folding and binding. *Protein Sci.*, 4:2499–2509.

Strynadka, N. C. J., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R. amd Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M., & James, M. N. G. (1996). Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. *Nature Struct. Biol.*, 3:233–239.

Totrov, M. & Abagyan, R. (1994). Detailed ab initio prediction of lysozyme-antibody complex with 1.6 Å accuracy. *Struc. Biol.*, 1:259–263.

Vakser, I. A. & Aflalo, C. (1994). Hydrophobic docking: A proposed enhancement to molecular recognition techniques. *Proteins: Struct. Funct. Genet.*, 20:320–329.

Appendix A

Methods: the BOXSEARCH docking program

This appendix presents a detailed description of the BOXSEARCH docking program, that has been developed by Trevor Hart in Randy Read's laboratory (Hart & Read, 1992; Hart & Read, 1994). Much of my own research has centred around the application of this tool to problems related to biomolecular association. The original published report of BOXSEARCH was descriptively titled "A Multiple-Start Monte Carlo Docking Method" - the method involves many cycles of a Monte Carlo-based docking algorithm.

Section A.1 is essentially an abbreviated and modified reprisal of descriptions of the underlying principles and procedures employed in BOXSEARCH, that were first presented elsewhere (Hart & Read, 1992; Hart & Read, 1994). Section A.2, offers a practical complement to the theoretical description, and is based largely on my own experiences using BOXSEARCH.

A.1 BOXSEARCH theory

Monte Carlo simulated annealing is an optimization method particularly well-suited to problems involving many local minima (Kirkpatrick et al., 1983), and therefore is of great utility in simulating protein folding and molecular interactions. Whereas standard energy minimization of a bimolecular complex will lead to convergence on the nearest local minimum, Monte Carlo simulated annealing may surmount significant energy barriers before converging on what may be a more favorable minimum.

According to statistical mechanics, the bulk properties of a system can be expressed as the weighted average of all of the states of the system. The weight accorded each state depends on the energy of the state (E_s) and the temperature of the system (T), and is described by the Boltzmann factor, $e^{E_s/kT}$. It is impractical to systematically calculate the average of a complex multi-state system. The Monte Carlo method overcomes this difficulty by performing a random sampling (of the system) that favors the states with the highest weight, and which therefore make the most significant contributions to the average.

In the Metropolis implementation of the Monte Carlo method (Metropolis et al., 1953), we start with a system at temperature T, and in the state s_i. A new state s' is generated by making a small change in s_i (in rigid-body docking this would be a small translation and/or rotation of the probe molecule) and the difference in energy ΔE of the two states is calculated. The new state for the system (s_{i+1}) is determined by the following rule: if $\Delta E < 0$ then $(s_{i+1}) = s'$; if $\Delta E > 0$, take s' as the new state s_{i+1} according to the probability distribution $p = e^{E/kT}$, otherwise take $s_{i+1} = s_i$. The second case is achieved by generating a random number $0 \leq r < 1$ and accepting the new state s' if $r < e^{E/kT}$. The new state will probably be accepted if the energy difference ΔE is small relative to kT, and will probably not be accepted if ΔE is large relative to kT. Following equilibration of the system, this procedure gives rise to a sequence of states s_i that are statistically distributed according to the

Boltzmann distribution: the average of this set is therefore a Boltzmann-weighted average.

Simulated annealing is incorporated into this process by commencing each Monte Carlo "run" at high temperature, and progressively lowering the temperature as the run progresses. Therefore, in the initial stages of the run the energy of the system will be relatively high, and energy barriers between adjacent minima will be easily traversed during the Monte Carlo procedure. As the system is cooled it will "freeze" into a local minimum. By performing many such (relatively short) Monte Carlo runs in a docking simulation, each starting from a different randomly chosen state, BOXSEARCH can effectively sample all of the important low energy states accessible to the system. The search space (*i.e.* the number of accessible states) is typically limited according to experimental information, such as knowledge of the catalytic residues of an enzyme (discussed in Chapter 2).

Scoring or ranking of dockings throughout a BOXSEARCH docking run is based on a pairwise atom interaction energy calculation that includes van der Waals and electrostatic terms (equation 1).

$$E_{interaction} = E_{vdW} + E_{elec} \tag{A.1}$$

Partial charges are combined into charge groups, and a distance cutoff (in most cases 8Å; employed to speed up the calculation) is applied to charge groups to avoid unrealistic cutoff artifacts. The van der Waals parameters and partial atomic charges are derived from the work of Hagler (Hagler et al., 1974; Lifson et al., 1979; Hagler et al., 1979a; Hagler et al., 1979b) and were obtained from John Moult. A united atom treatment is used, so that only polar hydrogen atoms are explicitly described. To compensate for the lack of consideration of solvent effects, the neutral charge group procedure of GROMOS is used (Aqvist et al., 1985). This has the effect of scaling down charge-charge interactions, which are over-estimated in the absence of solvent, while leaving other electrostatic interactions unaffected. In the case of non-protein atoms, such as those of the NAD and carbohydrate molecules employed in my research, charges were derived from the DISCOVER charge library (Biosym Technologies Limited, San Diego). The programs INSIGHTII and DISCOVER (Biosym Technologies Limited, San Diego) are used for energy minimization and general structure manipulation and visualization.

The initial phase of a BOXSEARCH docking run involves what is essentially simulated annealing with a shape-based score function. A grid (the *floating grid*) is constructed around and throughout the target molecule. Each grid point within the target is described by its distance to the nearest grid point outside of the target, and each grid point outside the target is given a value of 0. A score function that represents the distance of the probe to the surface of the target is used to float the probe to a position near the surface of the target. This is achieved by summing over all of the probe-heavy-atom/nearest-grid-point pairs, and minimizing the score function, using simulating annealing, until it either falls below a user-specified value (accepted) or exceeds a user-specified number of simulated annealing steps (rejected). This procedure is termed the *floating method.* When a docking is accepted after the floating method, the energy-based simulated annealing (described above) schedule is invoked. If the intermolecular interaction energy falls below the user-specified cutoff at the end of this procedure, the docking is accepted and written to output.

A.2 BOXSEARCH practice

A description of the theory behind the docking algorithm employed in the BOXSEARCH docking program was given in the preceding section. To complement this, and to give a more complete and detailed description of actually performing a docking simulation with this method, a general description (not describing a particular molecular system) is given here. The Methods sections of the following chapters each give a brief reiteration of this description, and supply details specific to the molecular systems being considered.

A.2.1 Probe and target preparation

BOXSEARCH uses a united atom approach, so only polar hydrogens bonded to heteroatoms (N, O, S) are explicitly described. Of these, the hydrogens of the sidechains of Lys, Tyr, Ser, Thr, and Cys are not fixed by stereochemical constraints, and therefore must be positioned by the user. His sidechains are initially protonated at NE2, and this is accepted unless examination of the structure indicates that possible *intra*molecular interactions favor protonation at ND1.

To begin with, all waters are deleted from the structures. Any protein sidechain substitutions are performed at this point, using INSIGHTII (Biosym Technologies Limited, San Diego). This same suite of programs is used to add and position (according to standard stereochemical rules) *all* hydrogens of the molecules being considered. For proteins, we then use the NETWORK program (Bass et al., 1992) to reposition mobile polar hydrogens so as to maximize *intra*molecular hydrogen bonding. This program also considers each of the two protonation sites of the His sidechain (see above). The output structure is then imported back into INSIGHTII, and prepared for energy minimization with the CVFF potential function of DISCOVER. All heavy atoms of the structure (protein or non-protein) are fixed, and *all* hydrogens (not just mobile polar hydrogens) are allowed to move during a maximum of 200 steps of steepest descents minimization, followed by a maximum of 200 steps of conjugate gradients minimization. If any protein sidechains were substituted (see above), the altered sidechains are allowed to move during the two rounds of post-NETWORK energy minimization. At this point, all non-polar hydrogens are removed from the molecules, and the atoms are re-named so as to be consistent with BOXSEARCH and related programs.

A.2.2 The floating grid

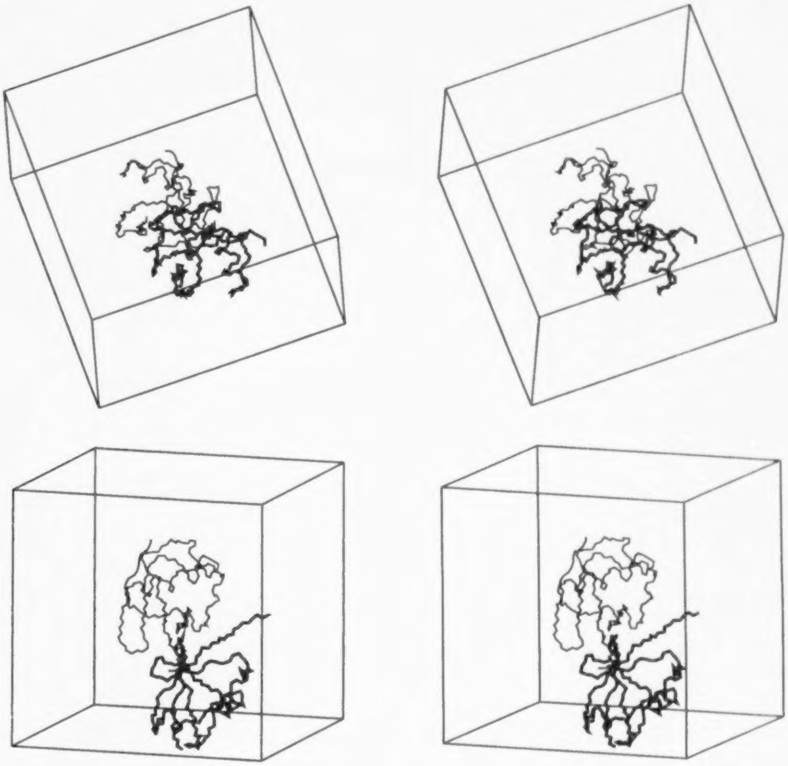


Figure A.1: Spatial extents of a BOXSEARCH simulation. The protein-protein system used for example purposes here is the diubiquitin system (see Chapter 2). The box represents the extents of the search space accessible to the probe in this particular BOXSEARCH simulation. The target protein (thick lines) is partly excluded by the box, and the probe (thin lines) is shown in its reference (correct) configuration.

Prior to doing any docking with BOXSEARCH, the floating grid (Section A.1) must be generated. The centre and dimensions of the space to be searched are chosen, typically limiting the extent of the search to some fraction of the target protein's surface. This choice is based upon either the known structure of the complex being docked, or indirect evidence, such as the location of catalytically essential residues, or residues that seem to be directly involved in the binding interaction of interest. For docking involving a relatively small probe, the entire target surface might be used. The extents of the chosen volume, relative to the docking target, can be visually inspected for correct placement with the INSIGHTII program (see Figure A.1). The final grid file is used as input for BOXSEARCH.

During the initial floating stage of each docking run, a docking will be rejected after a user-specified number of Monte Carlo steps if the *floating score* (related to the degree of overlap between the probe and target) is still exceeded at this point. The maximum number of floating steps and the floating score cutoff must both be specified.

During simulated annealing BOXSEARCH uses an annealing schedule that involves a gradual decrease in system energy and the size of the maximum rigid-body rotations and translations. A standard annealing schedule has been used for several different systems (Chapter 2). Prior to each energy calculation the floating contact score is calculated to determine the extent of overlap resulting from the last Monte Carlo step. This floating cutoff can be different from that specified for the floating stage of each docking run. If this cutoff is exceeded then the probe is automatically assigned a prohibitive interaction energy and this new state will be rejected, without performing a full energy calculation. This extra step results in a significant saving of what would otherwise be wasted calculation time.

A.2.3 Residue and charge group libraries

The user specifies coordinate files describing the probe and target molecules, and the relevant floating grid file. To compile all of the additional information needed to fully describe the docking simulation the program refers to library files. These files provide descriptions of atom type, partial atomic charge, and charge group composition for all of the atoms of the probe and target molecules. All of the residues in these two molecules must be accurately represented in the residue and charge group library files. For proteins, the library file modifications required for a new system are usually minor, in some cases requiring the inclusion of one or two new residues (*eg.* a truncated sidechain or an unnatural amino acid). If the probe is not protein then construction of the appropriate library files is required. For the non-proteinaceous molecules studied so far, we have been able to proceed from the charges supplied by the CVFF or AMBER (Weiner et al., 1984) potential functions of the DISCOVER program (Biosym Technologies Limited, San Diego). These were then scaled down to allow for neutral charge groupings, similar to the charge scalings employed in the BOXSEARCH peptide libraries.

A.2.4 Preliminary docking simulations

When all of the preparations described above have been completed, the elements of a BOXSEARCH docking simulation are ready. At this point several small simulations are run to establish a reasonable energy cutoff. Dockings which have interaction energies below this cutoff will be written to output, so a big simulation with a permissive cutoff could lead to a huge output file. For a new system it is also advisable to run several trials with different floating cutoffs, to ensure that dockings of interest are not being excluded by a particular parameter value. When suitable values for all of the user-specified parameters have been determined, larger simulations can then be performed.

A.2.5 Analysis of docking output

At the end of a BOXSEARCH simulation, the user is left with output that consists of a large coordinate file, with the interaction energy of each docking followed by its coordinates. Much of the analysis that will ultimately be performed on these results is specific to the system being studied and the problem being addressed. From my research, Chapter 2 provides the most extensive discussion of a variety of analysis techniques. The paper by Shoichet & Kuntz (1991) is also instructive regarding this aspect of docking. Certain general procedures will apply in the initial analysis of docking results for most systems, and these are outlined here.

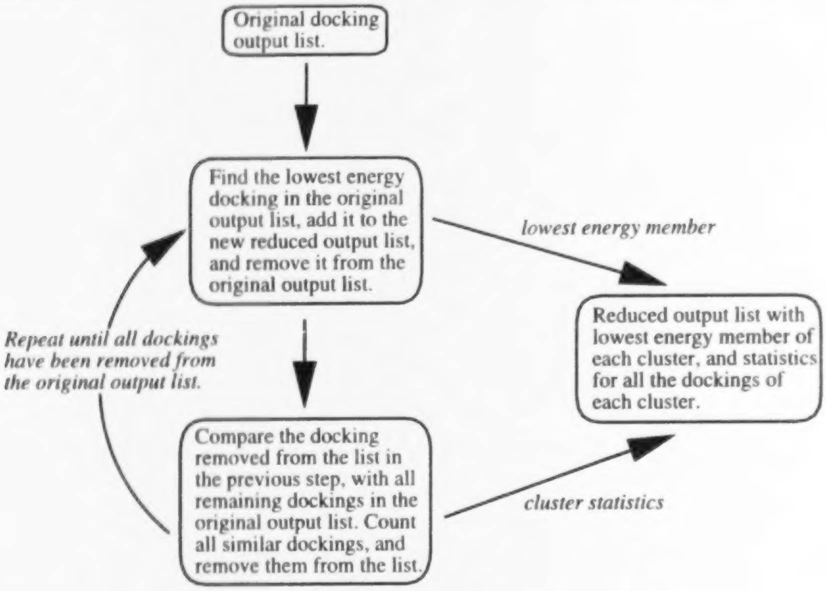

Original docking output list.
Find the lowest energy docking in the original output list, add it to the new reduced output list, and remove it from the original output list.
lowest energy member
Reduced output list with lowest energy member of each cluster, and statistics for all the dockings of each cluster.
Repeat until all dockings have been removed from the original output list.
Compare the docking removed from the list in the previous step, with all remaining dockings in the original output list. Count all similar dockings, and remove them from the list.
cluster statistics

Figure A.2: Cluster analysis of original docking output list. This procedure yields a reduced output list containing only the lowest energy representative of each distance-based cluster or family of dockings.

For all docking systems the first thing to do with the output is remove identical or similar dockings. We generally use an RMS distance criteria of 2Å for all heavy atoms, with a clustering program developed by Trevor Hart (Figure A.2). This procedure yields a reduced output list of distinct dockings, and lists statistics for each cluster such as the number of members of that cluster and the average energy for the cluster.

Depending on the docking system, there may or may not be a known correct (reference) answer. For the systems used for testing new docking methods a reference structure is always available. Typically this will be the experimentally determined structure of the complex, or the structure of one or both of the native components superimposed onto the appropriate halve(s) of the complex (the docking systems described in Chapter 2 provide three different examples of these types of reference complexes). Alternatively, a reference complex may be modeled, if fairly exact intramolecular distance constraints are available (these can be derived from a variety of experimental results - examples can be found in the work presented in Chapters 2 and 4). Whatever the origin of the model complex, prior to comparing it to dockings it is first minimized with the BOXSEARCH potential function. This will ensure that the reference complex represents an energy minimum accessible to the docking system during the simulation. In the earlier stages of my research this minimization was performed with a modified version of BOXSEARCH, with the annealing schedule shown in Chapter 2. More recently I have used a rigid-body conjugate gradients minimizer (unpublished program of Trevor Hart). The optimized model complex then becomes the ideal docking, and is compared to either the original or the clustered docking output. The best results that can be hoped for is that dockings similar (within 2Å RMS) to the reference have the lowest interaction energy, and are separated from more distant dockings by a significant energy difference. Several examples of such ideal results are shown in the scatter plots in Chapters 2 and 3.

In many applications of docking we seek to predict the structure of a complex for which we do not have a clear model. Such *ab initio* prediction of complexes is one of the ultimate goals of docking. The information that is available may be in the form of distance constraints derived from experiment (*eg.* the examples of Stoddard & Koshland (1992) and Yamamoto *et al.* (1994), discussed in the final section of the Introduction). We may know that a particular covalent intermediate exists transiently, or that a covalent product is formed (*eg.* Chapter 2, although an exact model was available), or that mutagenesis of certain amino acids implies their involvement, *eg.* Chapter 4), or lack thereof, (*eg.* Chapter 2, although an exact model was available), in the intermolecular interaction being studied. In these cases the reliability of the model will depend on the number and exactness of the distance constraints available, and how well the final model (or models) satisfies those constraints.

A.3 References

Aqvist, J., van Gunsteren, W. F., Leijonmarck, M., & Tapia, O. (1985). A molecular dynamics study of the C-terminal fragment of the L7/L12 ribosomal protein: Secondary structure motion in a 150 picosecond trajectory. *J. Mol. Biol.*, 183:461–477.

Bass, M. B., Hopkins, D. F., Jaquysh, W. A. N., & Ornstein, R. L. (1992). A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. *Proteins: Struct, Funct, Genet*, 12:266–277.

Hagler, A., Lifson, S., & Dauber, P. (1979a). Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 2. A benchmark for the objective comparison of alternative force fields. *J. Am. Chem. Soc.*, 101:5122–5130.

Hagler, A. T., Dauber, P., & Lifson, S. (1979b). Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 3. The CO···HO hydrogen bond and the analysis of the energetics and packing of carboxylic acids. *J. Am. Chem. Soc.*, 101:5131–5141.

Hagler, A. T., Huler, E., & Lifson, S. (1974). Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.*, 96:5319–5327.

Hart, T. N. & Read, R. J. (1992). A multiple-start Monte Carlo docking method. *Proteins: Struct, Funct, Genet*, 13:206–222.

Hart, T. N. & Read, R. J. (1994). Multiple-start Monte Carlo docking of flexible ligands. In Merz, K. M., J. & LeGrand, S. M., editors, *The Protein Folding Problem and Tertiary Structure Prediction*, pages 71–108. Birkhäuser, Boston.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

Lifson, S., Hagler, A. T., & Dauber, P. (1979). Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 1. Carboxylic acids, amides, and the CO···H - hydrogen bonds. *J. Am. Chem. Soc.*, 101:5111–5121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. J., & Weiner, P. (1984). A new force-field program for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784.

Appendix B

Flexibility in automated docking

To date, most automated docking methods have used completely rigid molecules. Protein-protein complexes have routinely been reconstructed with these rigid-body methods [reviewed in (Cherfils & Janin, 1993; Hart & Read, 1994; Lybrand, 1995)], and, in many test cases, reasonable models of the complexes have been achieved using the uncomplexed structure of one or both of the proteins [reviewed in (Cherfils & Janin, 1993; Hart & Read, 1994; Lybrand, 1995)]. In most of these studies the score function(s) used do not clearly and consistently distinguish the correct complex (the experimentally determined structure) from false positives [the report of Shoichet & Kuntz (1991) contains several examples and an excellent discussion of this problem]. In many cases the false positives differ dramatically from the correct configuration. A significant contributor to this deficiency is the inability of the interacting surfaces to optimize the complementarity of the interface, due to a lack of consideration of molecular flexibility. This approximation is unrealistic in most cases, and represents a less than satisfactory compromise.

Simple atom deletion (*eg.* sidechain truncation; (Shoichet & Kuntz, 1991); used and discussed in Chapter 2) and allowing for significant atomic overlap are two very crude, but in many cases effective, methods of approximating molecular flexibility.

Allowing for appreciable atomic overlap may be achieved by scaling down the van der Waals radii used to represent the molecules of interest, or by using a less steeply repulsive potential function than the standard 6-12 function (a "soft potential"). The latter approach was used, in conjunction with a very simplified residue representation, in the first reported automated docking method (Wodak & Janin, 1978), and most rigid-body docking methods do allow for some degree of atomic interpenetration. Both of these crude methods provide a half-measure of correction, at best. Allowing for overlap may overcome some prohibitive interactions that would be readily alleviated through minor conformational changes, but this does not represent an *optimization* of the interacting surfaces. Similarly, atom deletion will avoid bad clashes, but favorable interactions involving the deleted atoms will also be ignored.

The first attempt to approximate conformational flexibility in docking in a more realistic way was reported by Kuntz's group (DesJarlais et al., 1986). For two different protein-ligand systems they broke the small molecule ligands into two fragments and docked each pair separately to the relevant protein. Ligand dockings were reconstructed by checking for overlaps between chosen atoms of each fragment pair. For both test systems it was possible to reconstruct the observed complex structures from dockings of the fragments. However, the primitive score function used (shape matching) was only able to distinguish the correct docking in one case.

This work (DesJarlais et al., 1986) established the potential utility of fragment-based methods of *de novo* ligand design. Subsequently, such fragment-based design methods were implemented by other groups. Moon & Howe (1991) developed a method for constructing models of peptides bound to target proteins. The bound ligand is built up, one residue at a time, starting from an acetyl group terminus and drawing upon a library of amino acid conformers. The LUDI program (Bohm, 1992a; Bohm, 1992b) can suggest a large variety of natural and unnatural ligands based upon a simple set of rules and an extensive fragment library. This program has been implemented in the widely-used Biosym software suite (Biosym Technologies limited, San Diego).

Several recent reports describe fragment-based docking methods aimed specifically at structure-based drug design. Rotstein & Murcko (1993) developed the GroupBuild method, and reported tests using a very small library of fourteen small fragments. In this method a ligand core is "predocked" (by a variety of possible methods) to a completely rigid target protein, and then sequentially extended at user-defined (hydrogen) sites. New fragment-fragment torsions are extensively searched, and unfavorable values are rejected: first, according to a simple set of stereochemical rules, and, second, on the basis of probe-target clashes. Grid-based energy evaluation is used during the docking simulation, and some accounting for solvation effects is possible. Accepted structures are then subjected to a final round of more rigorous energy minimization with an external program.

Freer and co-workers have applied evolutionary programming to the problem of docking flexible ligands to a rigid target (Gehlaar et al., 1995b). Certain aspects of this approach are extensions of their previous work involving a Monte Carlo-based method of *de novo* ligand construction from a small but comprehensive set of fundamental atom types (Gehlaar et al., 1995a). In the two tests reported (another was mentioned but not described) a good correlation was observed between the calculated interaction energy and deviation from the experimental structure. What is particularly noteworthy about this work is the relative simplicity of the intermolecular terms of the potential function, which included only steric and hydrogen-bond terms. A discrete set of distance- and angle-based energy values was used to describe interatomic interactions. One notable limitation of this work is that in the test cases the probe molecules were restrained to a fairly small volume around the known binding site.

Such fragment-based approaches to approximating molecular flexibility are not directly applicable to protein-protein docking, and thus are limited to protein-small molecule docking. Since flexibility is only approximated *between* fragments, this approach also suffers from its reliance on a library of fragment conformers [a similar limitation for sidechain rotamers was noted and discussed by Schrauber et al. (1993)]. The ligand conformation in the lowest energy complex may not be composed of the available fragment conformers, even if several reasonable low energy conformers of each fragment are present in the library. Thus, the most favorable docking may be missed in the search. Strongly favoring this approach as a method of ligand design, however, is its usefuleness in designing a huge variety of possible ligands from a relatively small set of fragments. In this way it can serve as an "idea tool" for the medicinal chemist.

In the development of methods for representing conformational flexibility in protein-protein docking, several "conventions" seem to have been adopted. First, in most cases flexibility is considered for the probe, but not the target. This seems more acceptable for docking of small molecules to proteins than for protein-protein docking. In both types of systems this is a significant limitation. Given the significant increase in computational expense that molecular flexibility in docking represents, this may simply represent a first step towards the incorporation of both probe and target flexibility. Second, development has focussed on sidechain flexibility. Sidechain optimization may be sufficient for the prediction of many protein-protein interfaces, and the prediction of major backbone conformational changes represents a challenging problem (protein folding) in its own right. Here we review a few examples of protein-protein docking methods that have considered some degree of sidechain flexibility.

Goodsell & Olson (1990) made one of the first reports of automated docking involving conformational flexibility. Selected torsion angles of small probe molecules are considered as additional degrees of freedom to the standard translations and rotations of rigid-body docking. The target protein remains fixed throughout the

Monte Carlo-based docking search. To date this method has been applied to docking with small molecules [(Goodsell & Olson, 1990; Goodsell et al., 1993); discussed in the Introduction], and to a protein-protein system by using key peptides excised from the probe protein [(Stoddard & Koshland, 1992); discussed in the Introduction]. This method may be suitable for larger protein-protein applications, although such use has not yet been reported. Caflisch *et al.* (1992) have reported a similar method for protein-peptide docking. In both of these methods, the target protein remains fixed throughout the docking search.

Recently, methods that involve more extensive representations of conformational flexibility have been reported (Leach & Kuntz, 1994; Totrov & Abagyan, 1994; Stouten et al., 1995; and refs therein). These methods allow consideration of the conformational flexibility of both part or all the probe *and* target molecules. The degree of flexibility may also be scaled with respect to, for example, distance from a known binding site (Luty et al., 1995). For associations that do not involve significant backbone adjustments of the free molecules, these methods hold the promise of being able to locate *and* identify a complex at or near to the global minimum. The procedures of Leach (1994) and Totrov & Abagyan (1994) use fixed protein backbones, and rely on a search through a library of reasonable sidechain conformations [typically derived from a survey of high resolution structures (Janin et al., 1978; James & Sielecki, 1983; Ponder & Richards, 1987; Schrauber et al., 1993)] to optimize the conformations of interfacial sidechains [Schrauber et al. (1993) have discussed the limitations of the "rotamer library" method for sidechain optimization in folding and docking]. The method of Stouten et al. (1995) uses a grid representation of target residues distant from the binding site, and performs a full molecular dynamics simulation on the small molecule probe and active site residues of the target. This method seems adaptable to protein-protein docking, provided the docking search is limited to specified regions of the protein surfaces. Changes in conformational entropy that may occur upon complex formation are also considered in two of these methods (Leach, 1994; Totrov & Abagyan, 1994). The protein-protein study of Totrov & Abagyan (1994) is particularly encouraging. The coordinates of native lysozyme were docked to the HyHel5 antibody (coordinates of that bound to lysozyme), followed by biased-probability Monte Carlo optimization [refs in (Totrov & Abagyan, 1994)]. This lengthy procedure (500 CPU h total) yielded a solution 1.57Å RMS from the experimental structure, and separated by 19 kcal from the next most favorable solution. Incorporating the grid approximation of Stouten et al. (1995) into the method of Abagyan & Totrov (1994) would significantly reduce the computational expense of the latter method, with potentially small cost in accuracy (Luty et al., 1995). Also, hardware and algorithmic advances should make such computationally demanding procedures more generally applicable in the near future.

B.1 References

Bohm, H.-J. (1992a). The computer program ludi: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Design*, 6:61–78.

Bohm, H.-J. (1992b). Ludi: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Design*, 6:593–606.

Caflisch, A., Niederer, P., & Anliker, M. (1992). Monte Carlo docking of oligopeptides to proteins. *Proteins*, 13:223–230.

Cherfils, J. & Janin, J. (1993). Protein docking algorithms: simulating molecular recognition. *Curr. Op. Struc. Biol.*, 3:265–269.

DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D., & Venkataraghavan, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem*, 29:2149–2153.

Gehlaar, D. K., Moerder, K. E., Zichi, D., Sherman, C. J., Ogden, R. C., & Freer, S. T. (1995a). *De novo* design of enzyme inhibitors by Monte Carlo ligand generation. *J. Med. Chem.*, 38:466–472.

Gehlaar, D. K., Verkhivker, G. M., Retjo, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J., & Freer, S. T. (1995b). Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry & Biology*, 2:317–324.

Goodsell, D. S., Lauble, H., Stout, C. D., & Olson, A. J. (1993). Automated docking in crystallography: Analysis of the substrates of aconitase. *Proteins: Struct, Funct, Genet*, 17:1–10.

Goodsell, D. S. & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct, Funct, Genet*, 8:195–202.

Hart, T. N. & Read, R. J. (1994). Multiple-start Monte Carlo docking of flexible ligands. In Merz, K. M., Jr. & LeGrand, S. M., editors, *The Protein Folding Problem and Tertiary Structure Prediction*, pages 71–108. Birkhäuser, Boston.

James, M. N. G. & Sielecki, A. R. (1983). Structure and refinement of penicillopepsin at 1.8Å resolution. *J. Mol. Biol.*, 163:299–361.

Janin, J., Wodak, S., Levitt, M., & Maigret, B. (1978). Conformations of amino acid side-chains in proteins. *J. Mol. Biol.*, 125:357–386.

Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.*, 235:345–356.

Luty, B. A., Zacharias, M., Wasserman, Z. R., Stouten, P. F. W., Hodge, C. N., & McCammon, J. A. (1995). A molecular mechanics/grid method for evaluation of ligand-receptor interactions. *J. Comp. Chem.*, 16:454–464.

Lybrand, T. (1995). Ligand-protein docking and rational drug design. *Curr. Op. Struc. Biol.*, 5:224–228.

Moon, J. B. & Howe, W. J. (1991). Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *Proteins: Struct, Funct, Genet*, 11:314–328.

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193:775–791.

Rotstein, S. H. & Murcko, M. A. (1993). Groupbuild: A fragment-based method for *de novo* drug design. *J. Med. Chem.*, 36:1700–1710.

Schrauber, H., Eisenhaber, F., & Argos, P. (1993). Rotamers: To be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol.*, 230:592–612.

Shoichet, B. K. & Kuntz, I. D. (1991). Protein docking and complementarity. *J. Mol. Biol.*, 221:327–346.

Stoddard, B. L. & Koshland, D. E. J. (1992). Prediction of the structure of a receptor-protein complex using a binary docking method. *Nature*, 358:774–776.

Totrov, M. & Abagyan, R. (1994). Detailed ab initio prediction of lysozyme-antibody complex with 1.6 Å accuracy. *Struc. Biol.*, 1:259–263.

Wodak, S. J. & Janin, J. (1978). Computer analysis of protein-protein interactions. *J. Mol. Biol.*, 124:323–342.